ROBUST MULTIMODAL HAND- AND HEAD GESTURE RECOGNITION FOR CONTROLLING AUTOMOTIVE INFOTAINMENT SYSTEMS

Frank Althoff, Rudi Lindl and Leonhard Walchshäusl

BMW Group Research and Technology
Hanauerstr. 46, 80992 Munich, Germany
email: {frank.althoff, rudi.lindl, leonhard.walchshaeusl}@bmw.de

ABSTRACT

The use of gestures in automotive environments provides an intuitive addition to existing interaction styles for seamlessly controlling various infotainment applications like radio-tuner, cd-player and telephone. In this work, we describe a robust, context-specific approach for a video-based analysis of dynamic hand- and head gestures. The system, implemented in a BMW limousine, evaluates a continuous stream of infrared pictures using a combination of adapted preprocessing methods and a hierarchical, mainly rule based classification scheme. Currently, 17 different hand gestures and six different head gestures can be recognized in real-time on standard hardware. As a key-feature of the system, the active gesture vocabulary can be reduced with regard to the current operating context yielding more robust performance.

1. INTRODUCTION

When people talk among each other, information can be exchanged in a natural manner. Human beings are able to process several interfering perceptions at a high level of abstraction so that they can meet the demands of the prevailing situation. Inter-human communication is characterized by a high degree of expressiveness, comfort and robustness. Moreover, humans possess complex knowledge resources that are expanded permanent by continuous learning and adaptation processes in everyday life.

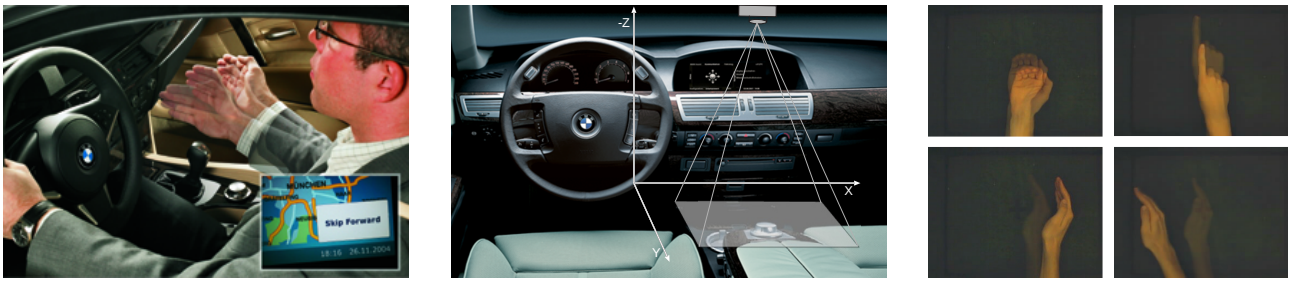
In contrast, exchanging information between humans and machines seems highly artificial. Many user interfaces show very poor usability, which is a result of growing functional complexity and mostly restriction to tactile input and visual output. Thus, the appropriate systems require extensive learning periods and adaptation to a high degree, which often increases the potential of errors and user frustration. To overcome these limitations, a promising approach is to develop more natural user interfaces that are modeled with regard to human communication skills.

Concerning human-machine interfaces the combination of various input and output resources like speech, gestures and tactile interaction is called multimodal interaction. In direct analogy to inter-human communication, multimodal interfaces have the potential to be more robust, since they integrate redundant information shared between the individual input modalities. Moreover, the user is free to choose among multiple interaction styles with regard to personal preferences.

In an automotive environment, the design of user interfaces has to cope with special requirements. The operation of driver information systems is a secondary task only that is subordinated to the control of the primary driving functions like steering, accelerating and braking. Seamless interaction by speech and gestures allows to use various in-car devices while keeping the eyes on the road. Gestures provide a comfortable addition to existing interaction styles. In direct comparison to speech interaction, gesture based input can even be used in noisy environments like driving in a convertible.

As a result of a longterm research cooperation between the Technical University of Munich and BMW Research and Technology, in this work we describe a robust and flexible system for the video-based analysis of dynamic hand- and head gestures that has been adapted to the individual needs of the driver and the specific in-car requirements. Moreover, the system is fully integrated in a multimodal architecture.

The paper is organized as follows. In section 2 we briefly explain the fundamental characteristics of gestures, describe relevant automotive usecase scenarios and review selected work in the field of automatic head- and hand gesture recognition. The overall system architecture is based on the classic image processing pipeline consisting of the two different stages spatial image segmentation (section 3) and gesture classification (section 4). This conventional process model has been extended by a spotting module that facilitates a fully automatic temporal segmentation of the continuous input stream. To increase the overall system performance, the entire parameter set can additionally be controlled by available context information of the user, the environment and the dialog situation (section 5). Finally, in section 6 we describe some experimental results of our system.



(a) (b) (c)

Fig. 1. (a) Skipping between audio tracks by hand gestures, (b) reference coordinate system for hand gestures with interaction area, (c) four gesture instances in motion (clockwise: down - XYSouth, up - XYNorth, left - XYEast), right - XYWest.

2. GESTURES

Depending on the specific research field, we can find various definitions of gestures. In his fundamental work, Kendon [1, 2] has explored in which way gestures are recognized by humans and, regarding a formal definition, identified the following aspects. Gestures correspond to a movement of individual limbs of the body and are used to communicate information. Moreover, the recognition of gestures can easily be done by humans as an unconscious process. Human beings can identify certain movements as gestures although they neither know the semantics nor the specific form of the gestures. With regard to a technical recognition system, gestures can be identified on the basis of a corresponding movement trajectory that is characterized by selected attributes like symmetry and temporal seclusion.

2.1. Application scenarios

In general, gestures facilitate a natural way to operate selected in-car devices. Thus, gestures can increase both comfort and driving safety since the eyes can focus on the road. A demonstration system has already been implemented in a BMW limousine. It gives the driver the possibility to perform a set of most frequently used actions.

The recognition of head gestures mostly concentrates on detecting shaking and nodding to communicate approval or rejection. Thus, head gestures expose their greatest potential as an intuitive alternative in any kind of yes/no decision of system initiated questions or option dialogs. As an example, incoming calls can be accepted or denied, new messages can be read, answered or deleted and help can be activated.

Hand gestures provide a seamless way to skip between individual cd-tracks or radio stations (see figure 1(a)) and to enable or disable audio sources. In addition, they can be used for shortcut functions, enabling the user to switch between different submenus of the infotainment system faster and more intuitive compared to standard button interactions. To increase both the usability and the robustness of the whole interface, the individual gestures can be interpreted in combination with spoken utterances and tactile interactions and vice versa. The gesture vocabulary has been derived from related usability studies [3]. Currently, the system is able to distinguish between 17 different hand gestures and six different head gestures. The four most important hand gestures are shown in figure 1(c).

2.2. Related work

Many research groups have contributed significant work in the field of gesture recognition. With regard to an automotive environment, Akyol [4] has developed a system called iGest, that can be used to control traffic information and email functions. Totally, 16 dynamic and six static gestures can be differentiated. The images are captured by an infrared camera that is attached to a active infrared lightning module. Due to the complex classification algorithms, only static gestures can be evaluated in real-time. Geiger [5] has presented an interesting alternative to a video-based system. In his work he used a field of infrared distance sensors to locate the hand and the head. The gesture vocabulary mainly consists of directional gestures to navigate within a menu structure and to control a music player. Although the sensor array does not achieve the resolution of a video-based image analysis, his system is highly robust and can get along with simple sensor hardware.

Concentrating on head gestures, Morimoto [6] has developed a system that is able to track movements in the facial plane by evaluating the temporal sequence image rotations. The parameters are processed by a dynamic vector quantization scheme to form the abstract input symbols of a discrete HMM which can differentiate between four different gestures (yes, no, maybe and hello). Based on the

IBM PupilCam technology, Davis [7] proposed a real-time approach for detecting user acknowledgements. Motion parameters are evaluated in a finite state machine which incorporates individual timing parameters. In an alternative approach, Tang [8] identifies relevant features in the optical flow and uses them as input for a neural network to classify the gestures. As an advantage the system is quite robust with regard to different background conditions.

3. SPATIAL SEGMENTATION

Detecting head- and hand postures in automotive environments requires illumination invariant techniques. Therefore, a near infrared imaging approach and a motion based entropy technique has been applied instead of conventional, mostly color based methods.

3.1. Adaptive Threshold

High reflectance of infrared radiation is characteristic of human skin (see figure 2(a)). Thus in the majority of cases the hand has shown to be the brightest object in the scene and can be found easily by a threshold operation. A static threshold is inapplicable for this purpose because of frequent illumination changes in the vehicle which are often caused by solar irradiation or driving through a tunnel. To overcome this problem we use a dynamic histogram based threshold in combination with near infrared imaging and active lighting.

This approach is based on the assumption that the current foreground object clearly differs in intensity from the background which results in a characteristic histogram behaviour. Correlative to the dominant grayvalue of the fore- and background two maxima and one corresponding minimum appear in the histogram. After smoothing this bimodal histogram with a Gaussian filter to reduce the effect of noise, this local minimum can be used as a dynamic threshold (see figure 2(c)). If the background consists of more than one object regarding its dominant brightness several minima will occur and can be used to discriminate these regions separately.

The presented technique gives promising results in a typical car environment at night and under low or diffuse lightning conditions. If the background consists of mostly plastic, wood or leather materials, this approach achieves high accuracy and is feasible of detecting 17 different hand gestures. Sceneries like hand postures where the area of the hand is too small to form a bimodal histogram or textiles such as cotton with nearly the same infrared reflectance coefficients as human skin are the main factors for potential misclassifications.

User studies have shown that a subset of five directional gestures are sufficient for controlling an audioplayer in an automotive environment. As the directional information associated with these gestures is more important than the accuracy of the segmentation, the hand detection process can be reduced to a more robust and plain motion based technique which is proposed in the next section.

3.2. Entropy Motion Segmentation

Motion detection is a fundamental task for many computer vision applications. In our application's environment the assumption can be raised that every motion within the gesture interaction area is caused by a moving hand. Thus we use a entropy based motion detection technique first presented in [9, 10] to detect moving objects in the scene. In this approach the intensity of every pixel is regarded as a state. Illumination changes, camera noise and moving objects are responsible for a pixel's state transition over time. Therefore the diversity of the state at each pixel can be used to characterize the intensity of motion at its position.

A temporal histogram is used to obtain a pixel's state distribution over time. To represent the relationship between one pixel and its neighbourhood both in time and in space this histogram is extended by the surrounding pixels $w \times w$ of the last L frames. As shown in figure 3 $w \times w \times L$ pixels are accumulated to build the temporal histogram of a pixel at location (i, j).

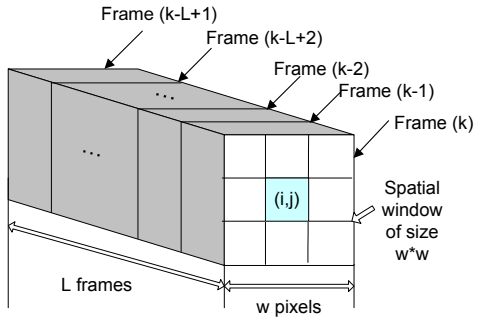


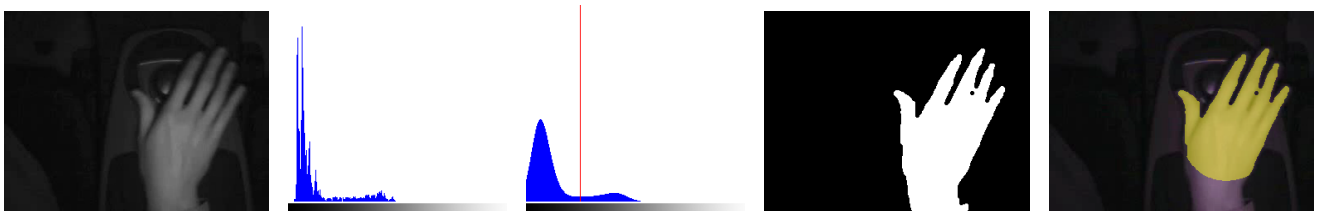
Fig. 3. Pixels used to form temporal histogram [9].

Computational effort can be reduced by quantizing the histogram into Q bins. After calculating the histogram $H(i,j)_q$, the probability density function $P(i,j)_q$ is derived by normalizing the histogram as following:

$$P(i,j)_q = \frac{H(i,j)_q}{w \times w \times L}, \qquad \sum_{q=1}^{Q} P(i,j)_q = 1 \quad (1)$$

Finally, the spatial temporal entropy $E(i,j)$ is defined by the following equation:

$$E(i,j) = -\sum_{q=1}^{Q} P(i,j)_q \cdot log(P(i,j)_q) \quad (2)$$



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

Fig. 2. Adaptive threshold segmentation: (a) input frame, (b) histogram of input frame, (c) local minimum in smoothed histogram, (d) binarized image with opening filter, (e) localized hand with truncated arm.

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

Fig. 4. Entropy motion segmentation: (a) IR camera image, (b) difference image, (c) entropy image, (d) binarized entropy image, (e) result after geometrical forearm filtering.

When motion occurs the histogram spreads wider and accordingly the entropy rises. The use of a spatial window causes edges in the image to result in comparable high entropy. To overcome this limitation the entropy is calculated on difference images (see figure 4(b)) instead of plain images as depicted in figure 4(a).

In order to suppress meaningless movements the entropy image has to be binarized (see figure 4(d)). Afterwards morphological operations remove areas of noise and clean up the remaining regions (see figure 4(d)). Finally, a forearm filtering process (see section 3.3) is applied on the region with the biggest area. The result as depicted in figure 4(e) is regarded as a moving hand and passed to the consecutive spotting process (see section 4.1).

3.3. Forearm Filtering

For the most part of spatial segmentation algorithms (especially colorbased and motionbased methods) the routines result in regions containing both the hand and the arm area. To filter the hand from the arm area a postprocessing step is indispensable. The following geometrical technique [11] is straightforward and has been chosen to ensure a computational feasible way.

The steps to be taken are as follows (see figure 5). Vertice C represents the centroid of the located hand-forearm-component. Vector $\vec{d}$ is determined by the orientation of the component and vertice C. The vectors $\vec{g}$ and $\vec{h}$ are put through C rotated with the angle $\omega = 45$ degrees from $\vec{d}$. Slicing g and h with the contour of the hand-arm-component results in the vertices G and H.

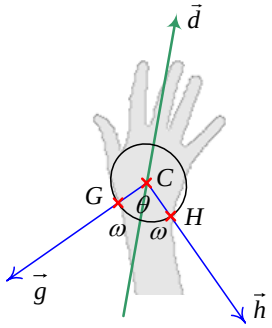


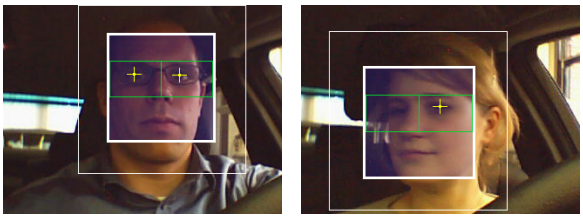
Fig. 5. Geometrical forearm filter

An ellipse sector through G and H with center C is defined unambiguously w.l.o.g. through the long axis $\overline{CG}$ and the short axis $\overline{CH}$. The boundary $\theta(\vec{CG}, \vec{CH})$ of the ellipse sector forms the cut surface between arm region and hand region.

A further movement of the arm region into the display

window, results in a strengthened translation of the centroid C towards the forehand area. To ensure nevertheless a proper filtering process, the operation is repeated until the ratio of the rotated bounding box of the resulting hand converges to $\gamma = 0.3$.

3.4. Formbased Headlocalization



(a) (b)

Fig. 6. Formbased head and eye segmentation. Head search area (red box), eye search area (green box), found head (blue box), found eye (yellow cross). (a) frontal view of the face, (b) shake gesture

The detection of the head is a common challenge in many applications like face recognition or pose estimation. In contrary to hand segmentation the appearance of the frontal face is only affected by rotational movements and facial expressions. Therefore a form-based segmentation algorithm [12] has been chosen to localize the head and to extract all relevant facial features. Two cascades of simple classificators have been trained to extract the face and eye positions. The trainingset consists of 3000 face respectively eye samples and 1000 negative background images. The initial head extraction is performed on the whole image. Further searching steps are limited to the last head position increased by an additional confidence area. Likewise the search region for the eyes is limited to the upper half of the extracted head region (see figure 6(a) and 6(b)). These restrictions of the extraction zones allow on the one hand image processing in real-time and on the other hand a more robust head tracking and feature extraction.

Typical head gestures like shaking and nodding are performed with periodical rotations of the head. These movements result in characteristic Al trajectories of the eye regions within the 2D image plane. If both eyes are visible the tracking reference point is set to the center between the two eye region. At certain head postures only one part of the face is visible to a frontal viewer and one eye is occluded for the most part. In this case the trajectory of the unoccluded eye-region is taken as reference point for the consecutive spotting process (see section 4.1).

4. CLASSIFICATION

4.1. Temporal Segmentation

Gesture spotting refers to the extraction of a meaningful temporal segment corresponding to gestures from continuous input streams that vary both in space and time. By using an automatic spotting module, the user is able to interact with the system without explicitly keeping the start and the end of the gestures in mind. Considering commonly known physiological gesture characteristics [1, 2], a set of rules can be deduced to distinguish meaningless movements from relevant gestures. Since all gestures are associated with movement, an appropriate motion indicator has to be introduced which shows possible gesture parts. The feature based indicator $M(t)$ is defined by

$$M(t) = \sqrt{(\Delta X)^2 + (\Delta Y)^2} \tag{3}$$

where ΔX and ΔY describe the discrete derivations of the position from the segmented hand and head, respectively (see figure 1(b) for reference coordinate system). To ignore noise and small meaningless movements a threshold T is introduced which forms the binary motion trigger $I(t)$.

$$I(t) = \begin{cases} 0 & M(t) \leq T \\ 1 & else \end{cases} \tag{4}$$

As motion is a inevitable but not sufficient criteria for a correct temporal segmentation of gestures, the following additional rules are introduced to minimize false detections (see figure 7). Every valid gesture g with start time t_b and end time t_e has to satisfy the following rules:

- *Rule 1 (Intergesture distance)*
  To avoid fast consecutive gesture executions, an intergesture duration $c_i = 1s$ is used to limit the time between two succeeding gestures g_{n-1} and g_n.

$$t_{b,n} - t_{e,n-1} \stackrel{!}{\geq} c_i \tag{5}$$

- *Rule 2 (Start-criteria)*
  The beginning frames $c_b = 3$ of a valid gesture g_n have to consist of motion. Every gesture has to be performed within a circular interaction area A_b with corresponding centerposition P_{start} and radius T_{start}.

$$\sum_{t=t_b}^{t_b+c_b-1} I(t) \stackrel{!}{=} c_b, \qquad \overline{P(t_b)P_{start}} \stackrel{!}{\leq} T_{start} \tag{6}$$

- *Rule 3 (End-criteria)*
  To overcome short resting points or parts with low motion the end of a gesture g_n is indicated by $c_e = 4$ consecutive frames with no motion. As gestures have

shown a symmetric behaviour the distance between the spatial start position $P(t_b)$ and the end position $P(t_e)$ has to be less than $T_{dist} = 60$.

$$\sum_{t=t_e-c_e}^{t_e-1} I(t) \stackrel{!}{=} 0, \qquad \overline{P(t_b)P(t_e)} \stackrel{!}{\leq} T_{dist} \quad (7)$$

- *Rule 4 (Maximum and minimum gesture length)*
  The maximum and minimum gesture durations $c_{max} = 3s$ and $c_{min} = 0.5s$ reject short noise and unusual long movements.

$$t_e - t_b \stackrel{!}{\geq} c_{min}, \qquad t_e - t_b \stackrel{!}{\leq} c_{max} \quad (8)$$

- *Rule 5 (Bad frame tolerance)*
  A bad frame tolerance $T_{bad} = 10$ is introduced to tolerate short segmentation dropouts.

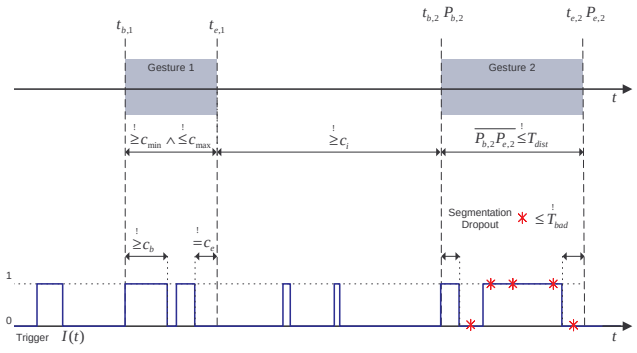


Fig. 7. Spotting parameter

All movements compliant to these rules are passed to the classification module, which performs the final recognition task.

A hierarchical rule based scheme [13] for dynamic gesture recognition has been chosen to ensure context integration and high performance. At the first phase of classification, the gesture to be recognized is assigned to one of two classes based on its dominant feature trajectories. Dominant features are features capturing the majority part of information conveyed by the gesture. The final decision is made up in the second phase by a class specific chart analysis considering the location, quantity and distribution of function minima and maxima in the dominant feature trajectories. As each class consists of two or three gestures, the probability of false classifications can significantly be reduced resulting in a more robust recognition process.

- **Phase 1 (Initial classification based on dominant trajectories)**

  Having smoothed the function with a median filter followed by a Gauss filter to eliminate outliers (compare figure 8(c) and 8(d)), the maximum amplitudes δ_I of the X- and Y-trajectories $(X(t), Y(t))$ of the gesture are determined:

$$\delta_I = \max(I(t)) - \min(I(t)) \text{ with } I \in \{X, Y\} \quad (9)$$

The following decision flow determines the reduced gesture set considering the dominant trajectories.

Set 1 (XYWest, XYEast, XWipe): Gesture-Set 1 contains the gestures along the X-axis and is chosen, if the X-axis amplitude exceeds a threshold θ_1 and the Y-axis–X-axis ratio is sufficient (see figure 8(b) for example):

$$(\delta_X \geq \theta_1) \wedge (\delta_Y < \frac{\delta_X}{2}) \quad (10)$$

Set 2 (XYNorth, XYSouth): Analog to Set 1, Set 2 contains the gestures along the Y-Axis and is chosen, if the Y-axis amplitude exceeds a threshold θ_2 and the X-axis–Y-axis ratio is sufficient (see figure 8(d) for example):

$$(\delta_Y \geq \theta_2) \wedge (\delta_X < \frac{\delta_Y}{2}) \quad (11)$$

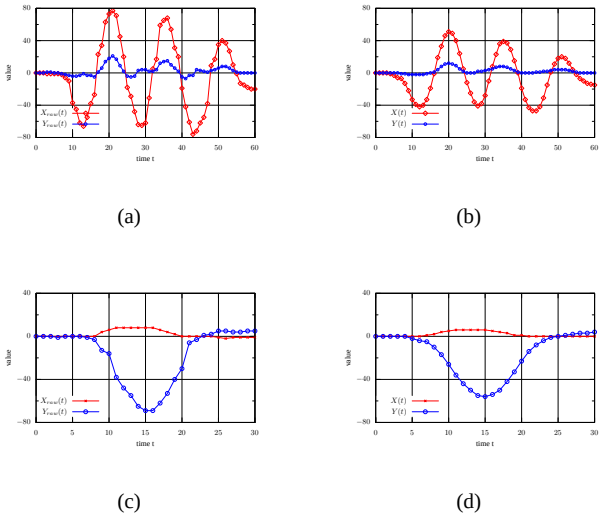


Fig. 8. (a) Trajectory of a XWipe gesture where $X(t)$ is the dominant trajectory and (c) of a XYNorth gesture where $Y(t)$ is dominant; (a,c) Raw input, (b,d) median and Gauss filtered data.

- **Phase 2 (Trajectory based classification)**

  After assigning the gesture to one of the two sets by means of phase 1, a detailed function analysis concludes the recognition process. Therefore, the dominant function trajectories are split-up into subfunctions at their extrema values. Decisive for the final

gesture determination is the succession of the subfunction gradients, the extrema count and extrema distribution.

4.2. HMM-based Classification

Hidden Markov Models (HMM) have originally been applied in the field of automatic speech recognition. In the last years they have successfully been used for other dynamical classification tasks, too, such as gesture recognition [14, 15]. To reduce the amount of information provided by the image sequences, the gestures are limited to their relevant data, consisting of the trajectory, velocity and hand form of the gesture. These samples are used to train a stochastic model for every gesture. In the recognition phase an output score is calculated for each model from the active input sequence, giving the probability that the corresponding model generates the underlying gesture. The model with the highest output score represents the recognized gesture. See [16] for an in-depth discussion of HMM.

Topology

In contrast to an ergodic HMM where the topology of the model is a fully connected graph, the left-right HMM λ (see figure 9) has the following restriction. From the current state q_i only a certain amount k of successor states $q_{i+j|0\leq j\leq k}$ are reachable. This topology has been chosen, because it is perfectly suitable to model temporal processes like gestures.

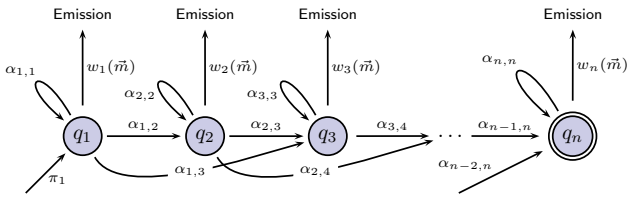


Fig. 9. Left-right topology ($k = 2$) of a semicontinuous Hidden Markov Model (sHMM).

Training

A gesture is converted to a sequence of feature vectors $\vec{m}_1 \dots \vec{m}_s$. A single vector $\vec{m}_t$ contains the change in position (ΔX and ΔY) and the Hu moments ($H_1 - H_6$) [17] of the hand area at the current time t. For head gestures only the position features are relevant.

$$\vec{m}_i^{hand} = \begin{pmatrix} \Delta X \\ \Delta Y \\ H_1 \\ \vdots \\ H_6 \end{pmatrix} \quad \vec{m}_i^{head} = \begin{pmatrix} \Delta X \\ \Delta Y \end{pmatrix} \quad (12)$$

For every gesture g a left-right sHMM λ_g is trained with $k = 2$ and $n = \frac{\overline{s}_g}{2}$ using the Viterbi-algorithm. A training stock of 250 samples per gesture showed to be adequate (see also figure 11).

Recognition

In the classification step every trained model λ_g is fed with the incoming feature sequence $\vec{m}_1 \dots \vec{m}_n$ from the temporal segmentation module. A gesture is perceived, if the generation probability density $P(\lambda_g)$ of its model comes first and exceeds a threshold θ_λ ($\theta_\lambda = -120$).

5. CONTEXT INTEGRATION

Especially in automotive environments a variety of sensor information is produced, which interpretation results in context information (see figure 10). The terms context information and context knowledge, which are used synonymic in this paper, are composed of the following contents [18]:

- *System context* creates information about the availability of single modules or the current system state like the velocity or the current state of the user interface, etc.
- *Environment context* combines all information, which can be directly deduced out of the surrounding environment (Lighting conditions, soundscape, etc.).
- *User context* provides information about the user and its behaviour. For example user preferences, usage history, gender, age, etc.

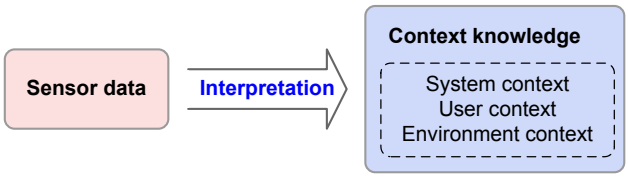


Fig. 10. Context knowledge is generated by the interpretation of sensor data and meta information that is based on sequences of user interactions.

5.1. Temporal Segmentation

As the temporal segmentation only consists of few rules and certain parameters, context integration is limited to the introduction of new rules or the adaption of existing parameters. In particular lower false positive rates and better detection rates can be expected as more information increases the selectivity of the process. Nevertheless some of the rules are

too general to be modified in a reasonable way. Only the threshold of the motion indicator (see formula 4) and the minimal and maximal duration of gestures are utilized and explained in the following.

The quality of the motion indicator can be increased if context information is utilized to limit the influence of specific moving directions ΔX and ΔY. The false positive rate is decreased for example if the system context forbids gestures within the y-axis. This approach, however implies that every gesture can be separated in groups regarding their dominant moving direction. The gesture vocabulary proposed in this work was chosen considering this aspects.

User studies have shown that gestures have distinct execution durations. Thus rule 4 (see section 4.1) has been introduced to reflect these characteristics. If no context information is available the minimum and maximum gesture duration has to be suitable for all gestures in the vocabulary v_a. A more precise model of the gesture execution is achieved, if these boundaries are adapted depending on context or gesture information. Accordingly the false positive rate is expected to be reduced.

$$c_{min} = \min_{\forall\, g \,\in\, v_a} \left[\frac{d_{min,g}}{P_c(g)} \right] \tag{13}$$

$$c_{max} = \max_{\forall\, g \,\in\, v_a} \left[d_{max,g} \cdot P_c(g) \right] \tag{14}$$

$P_c(g)$ denotes the contentual probability that gesture g (element of the complete gesture vocabulary v_a) will be performed. Respectively $d_{min,g}$ and $d_{max,g}$ refer to the minimal and maximal execution duration of gesture g.

5.2. Context-Integration for Classification

In general, the integration of context information into the rulebased classification process is performed by context sensitive strenghtening or weakening of the feature trajectories. Moreover, gestures that are not in the current context state are discounted from the classification process. More precisely the actual context state c gives the likelihood $P_c(g)$ that a particular gesture g occurs. The context probability for any set $G \subset v_r$ of gestures results in

$$P_c(G) = \min_{\forall\, g \,\in\, G} P_c(g) \tag{15}$$

The composed context probability $P_c(G)$ is multiplied with the feature trajectories of formula (9).

$$\begin{aligned} \tilde{I}(t) &= P_c(G_I) \cdot I(t) \quad \text{with } I \in \{X, Y\} \\ G_X &= \{\mathsf{XYWest}, \mathsf{XYEast}, \mathsf{XWipe}\} \\ G_Y &= \{\mathsf{XYNorth}, \mathsf{XYSouth}\} \end{aligned}$$

These actions diminish the influence of noise in unlike gesture trajectories and reduce therefore the probability for false classifications.

The context integration for the sHMM-based classifier is performed similarly except that all generation probabilities $P(\lambda_g)$ have to be normalized before the context probability-based scaling. For the reason that these probability densities have a value range of $]-\infty;\infty]$ all generation probability have to be subtracted with the minimal value of λ_g to raise them to values ≥ 0 before they are scaled with the appropriate context probability $P_c(g)$.

$$(\forall g \in v_a) \quad P_c(\lambda_g) = \left[P(\lambda_g) - \min_{(\forall g \in v_a)} P(\lambda_g) \right] \cdot P_c(g) \tag{16}$$

6. EXPERIMENTAL RESULTS

In this section, we will briefly describe some essential experimental results for the evaluation of the individual system modules and the performance of the overall system as implemented in the BMW limousine. Thereby, the results are discussed both with and without the use of additional context knowledge.

6.1. Temporal Segmentation

In the following a manually spotted videostream of 400 gesture executions is used to determine the recognition and false positive rate of the spotting module. An automatically spotted gesture g is correctly detected if it is overlapping at least 90% and at most 120% with the manually tagged gesture. To simulate a natural user behaviour several test persons had to perform up to five incarnations of the requested gestures in an arbitrary order. In addition a high amount of non gesture movements is guaranteed, as every gesture is embedded in a typical driving task like gear shifting, adjusting the fan, etc.

Both the motion indicator and the suggested rules have shown to model the natural gesture behaviour very well. Detection rates of over 98% have been achieved. Most of the missed gestures are rejected because a wrong end point was detected. This happens if the measured motion drops below the motion indicator threshold and rule three is satisfied. In comparison to the manually determined start and ending points the automatically obtained results differ on the average by 1.7 respectively 2.4 frames. As the rulebased approach lacks knowledge about the used gesture vocabulary a high false positive rate of 4.8 false detections per minute was to be expected.

6.2. Classification modules

The evaluation of the classification was performed with a gesture set composed of the five most important instances (*left*, *right*, *forward*, *backward* and *wipe*). Both person-specific as well as person independent testings have been un-

dertaken. The rulebased and the HMM-based classification scheme achieve similar gesture recognition rates around 90% on the selected gesture set. The two algorithms differ mainly from each other in their robustness against spatial segmentation errors and the training effort. As shown in figure 11 the HMM-based person-specific approach requires at least 150 trainings samples to achieve acceptable classification performance. In contrast, the rulebased approach needs no explicit training pool. Every gesture corresponds to a hard coded rule. Only the parameters have to be adapted correlatively.

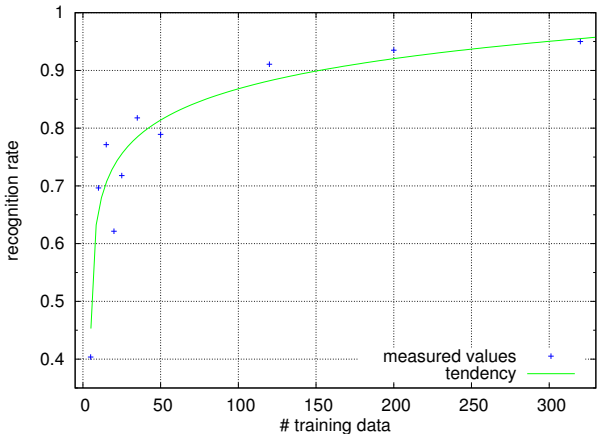


Fig. 11. Relation of recognition rate to training data.

Failures in the spatial segmentation process result in falsified shape and trajectory features. Especially the HMM-based classifier is affected by these erroneous modifications. The rulebased approach is able to compensate these flaws because of the initial median filtering which eliminates outliers and because of the hierarchical manner of the algorithm.

6.3. System performance

To obtain a robust, person-independent recognition performance of the individual modules, the system has been trained with 1530 hand gesture samples from six different users. The complete set v_a of 17 different gestures was equally distributed among these samples. Based on preliminary usability tests, a second set v_r of the five most frequently used gestures has been extracted. On the average, 86% (v_a) and 93% (v_r) of the gestures are classified correctly.

By using additional context information, e.g. input from other interaction modalities or reducing the meaningful gesture vocabulary due to the current dialog situation, recognition rates of nearly 97% can be achieved, and, what is more, the false acceptance rate can be decreased by 30%. Moreover, if person-specific data is used, e.g. in combination with dedicated keys associated with individual drivers, the system performance can additionally be increased.

The detailed evaluation of the head gesture module is subject to current work. First tests show highly promising results with average recognition rates slightly above the hand gesture module. In general, further usability testing has to evaluate the potential of gesture input with respect to various driving situations.

7. CONCLUSION

Compared to classical, mostly tactile interaction paradigms, gestures provide an interesting alternative for controlling selected in-car infotainment applications. Gestures are an important part of inter-human communication. Thus, the automatic recognition of gestures in an automotive environment can increase both the usability of complex driver information systems and driving safety since the eyes can be kept on the road. Head gestures show their greatest potential as an intuitive input form in any kind of yes/no decision of system initiated questions or option dialogs, e.g. accepting or denying an incoming call. Hand gestures provide a seamless way to skip between individual cd-tracks or radio stations and to navigate in a map. Emulating human behaviour, the individual gestures can be interpreted in combination with spoken utterances and tactile interactions. Moreover, the active gesture vocabulary can be reduced to meaningful gestures using additional context information. Concerning the recognition process, the system evaluates a continuous stream of infrared pictures using a combination of adapted preprocessing methods and a hierarchical, mainly rule based classification scheme. In general, 17 different hand and six different head gestures can be recognized using simple, state of the art hardware. Experimental results show that the automatic recognition of gestures contributes to the design of both effective and intuitive in-car user interfaces.

8. REFERENCES

[1] A. Kendon, "Current Issues in the Study of Gesture," *Anthropological Study of Human Movement*, vol. 5, no. 3, pp. 101–134, 1989.

[2] A. Kendon, "An Agenda for Gesture Studies," *Semiotic Review of Books*, vol. 7, no. 3, pp. 8–12, 1996.

[3] Frank Althoff, Gregor McGlaun, Manfred Lang, and Gerhard Rigoll, "Evaluating multimodal interaction patterns in various application scenarios," in *Proc. of Gesture Workshop 2003*, April 2003, Genua, Italien.

[4] U. Canzler S. Akyol, *GeKomm - Gestenbasierte Mensch-Maschine Kommunikation im Fahrzeug*,

Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 01 2000.

[5] M. Geiger, *Berührungslose Bedienung von Infotainment-Systemen im Fahrzeug*, Ph.D. thesis, TU-München, 2003.

[6] C. Morimoto et al., "Recognition of head gestures using hidden markov models," *Proc. of IEEE Int. Conf. on Pattern Recognition*, Vienna, Austria 1996.

[7] J. Davis et al., "A perceptual user interface for recognizing head gesture acknowledgements," *In WS on Perceptive User Interfaces (PUI 01)*, USA 2001.

[8] J. Tang et al., "A head gesture recognition algorithm," *In Proc. of the 3rd Int. Conf. on Multimodal Interface*, Beijing, China 2000.

[9] Chng Eng Siong Guo Jing and Deepu Rajan, "Foreground motion detection by difference-based spatial temporal entropy image," Tech. Rep., Nanyang Technological University, 2004.

[10] Hong-Jiang Zhang Yu-Fei Ma, "Detecting motion objects by spatio-temporal entropy," Tech. Rep., Microsoft Research, 2001.

[11] U. Bröckl-Fox, *Untersuchung neuer, gestenbasierter Methoden für die 3D Interaktion*, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 1995.

[12] Paul A. Viola and M. J. Jones, "Robust real-time face detection.," in *ICCV*, 2001, p. 747.

[13] James P. Mammen, Subhasis Chaudhuri, and Tushar Agarwal, "A Two Stage Scheme for Dynamic Hand Gesture Recognition," in *Proceedings of the National Conference on Communication (NCC 2002)*, 2002, pp. 35–39.

[14] Peter Morguet and Manfred Lang, "A universal hmm-based approach to image sequence classification," in *Proceedings of the 1997 International Conference on Image Processing (ICIP '97) 3-Volume Set-Volume 3*. 1997, p. 146, IEEE Computer Society.

[15] H. Lee and J. Kim, "An HMM-Based Threshold Model Approach for Gesture Recognition," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 10, pp. 961–973, October 1999.

[16] L. R. Rabiner, "A tutorial on hidden markov models and selected apllications in speech recognition," in *Readings in Speech Recognition*, A. Waibel and K.-F. Lee, Eds., pp. 267–296. Kaufmann, San Mateo, CA, 1990.

[17] M. K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions in Information Theory*, vol. IT-8, pp. 179–187, 1962.

[18] Manfred Lang, "FERMUS: Fehlerrobuste Multimodale Sprachdialoge," Tech. Rep., Lehrstuhl für Mensch-Maschine-Kommunikation, Juli 2003, Auftragsforschung für die BMW Group, die Daimler-Chrysler AG und die Siemens-VDO Automotive AG.