High Performance Real-Time Gesture Recognition Using Hidden Markov Models

Gerhard Rigoll, Andreas Kosmala, Stefan Eickeler

Gerhard-Mercator-University Duisburg
Department of Computer Science
Faculty of Electrical Engineering
Duisburg - Germany
e-mail: {rigoll,kosmala,eickeler}@fb9-ti.uni-duisburg.de

Abstract. An advanced real-time system for gesture recognition is presented, which is able to recognize complex dynamic gestures, such as "hand waving", "spin", "pointing", and "head moving". The recognition is based on global motion features, extracted from each difference image of the image sequence. The system uses Hidden Markov Models (HMMs) as statistical classifier. These HMMs are trained on a database of 24 isolated gestures, performed by 14 different people. With the use of global motion features, a recognition rate of 92.9% is achieved for a person and background independent recognition.

1 Introduction

The recognition of human gestures is a research area with increasing importance for multi-modal human computer interfaces. As reported in [1], different approaches for the recognition of gestures in video sequences are known.

A real-time recognition system for video sequences is presented in [2], where a superposition of the difference images is used for feature extraction. One of the first approaches with Hidden Markov Models (HMM) for the recognition of American sign language is presented in [3].

Gesture recognition is often done by means of e.g. colored gloves [3] or data gloves as described in [4]. The intend of the approach proposed here, is the use of dynamic pattern recognition methods, using only visual information without further aids. Furthermore, a recognition system should be able to recognize the gestures in person-independent mode. For the use in real-world applications it is necessary, that the system does not fail outside the laboratory environment. Another important condition for it's application is the use of a standard PC for the recognition. Up to now only the presented approach meets all these demands.

The system has some special features, in order to achieve high recognition rates for a 24 gesture recognition task in real-time:

- The feature extraction reduces the amount of data to 0.3 % of the original amount and describes the dynamics of the motion in the video sequence.
- A significant error reduction was achieved with the use of Hidden Markov Models instead of Neural Networks [5, 6].

- Further, the switch from discrete HMMs to continuous HMMs improved the robustness of the system in combination with the new feature extraction method.

2 Database and Recognition Task

For the training and the test of the system we recorded video sequences of 24 different isolated gestures.[1] The resolution of the video sequences was 96 × 72 gray-scale pixel and 16 frames per second. Each sequence consists of 50 frames, resulting in a sequence length of approximately three seconds. We recorded 24 gestures of 14 different people. This results in a database of 336 video sequences. The people performed the gestures spontaneously and were not trained on the gestures.

To get the maximum amount of training and test data we used the hold out method. All sequences of one person were removed from the complete set of 336 sequences, and the recognition system was trained with the remaining 312 samples. For the test we used the 24 sequences that were removed from the complete set before. This process was repeated for all people. Finally we calculated the average recognition rate over all people.

Figure 1 shows the different gestures in our database. The first 16 gestures are periodic gestures, that were repeated several times during the recording of three seconds. The last eight gestures were performed only once within the recording time.

3 The Recognition System

Figure 2 gives an overview of the recognition system. It contains three processing levels:

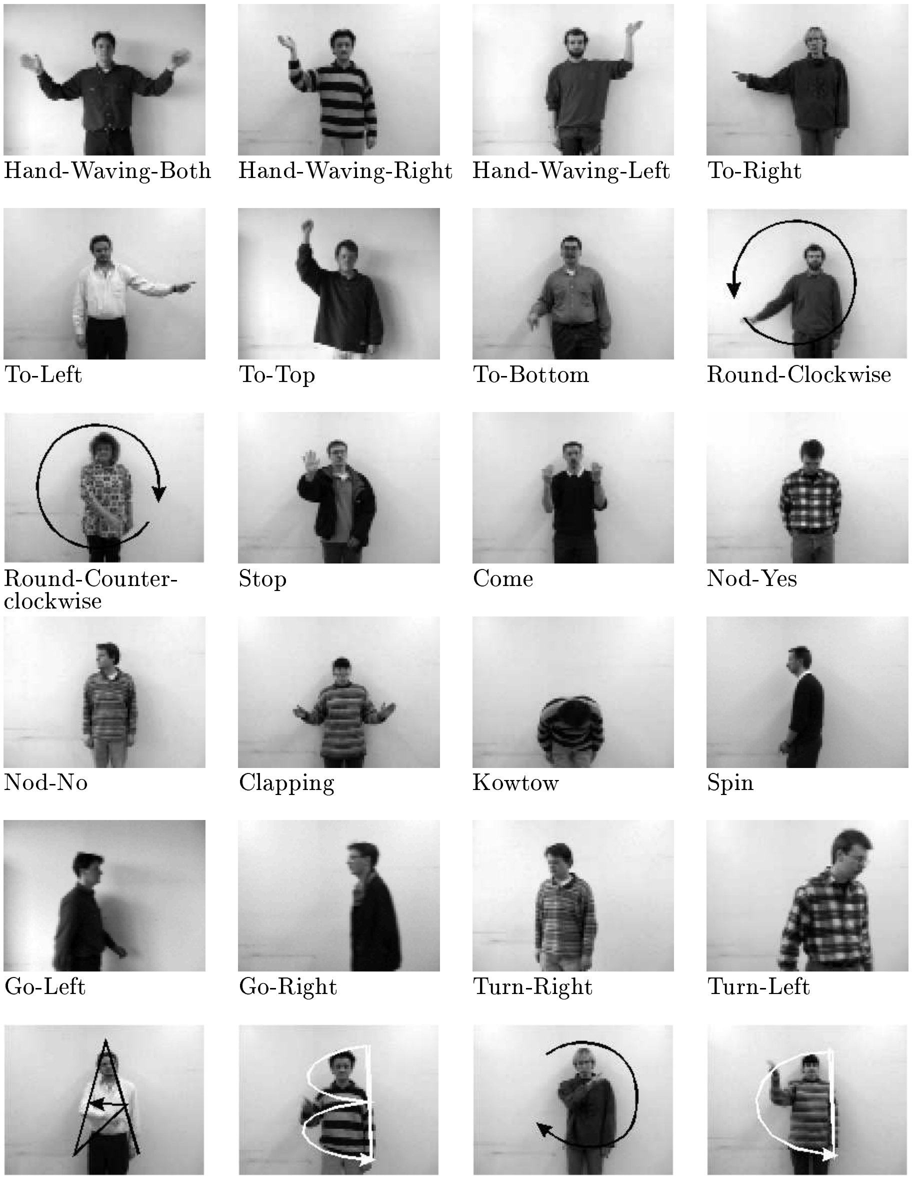
- preprocessing
- feature extraction
- statistical classification

Starting with the original image sequence of the gesture, the preprocessing calculates the difference image sequence. The feature extraction calculates a vector for each frame of the difference image sequence. A subsequent HMM based recognition module classifies the gestures, represented by the feature vector sequence. The output of the system is the recognized gesture.

3.1 Preprocessing

The preprocessing prepares the image sequence for the recognition by eliminating the background. As already shown in [5], the difference image sequence, has

[1] gesture samples are on the web-page http://www.fb9-ti.uni-duisburg.de/projekte/video/video.html (in German)



Hand-Waving-Both
Hand-Waving-Right
Hand-Waving-Left
To-Right
To-Left
To-Top
To-Bottom
Round-Clockwise
Round-Counter-clockwise
Stop
Come
Nod-Yes
Nod-No
Clapping
Kowtow
Spin
Go-Left
Go-Right
Turn-Right
Turn-Left
A
B
C
D

Fig. 1. Samples of the different gestures

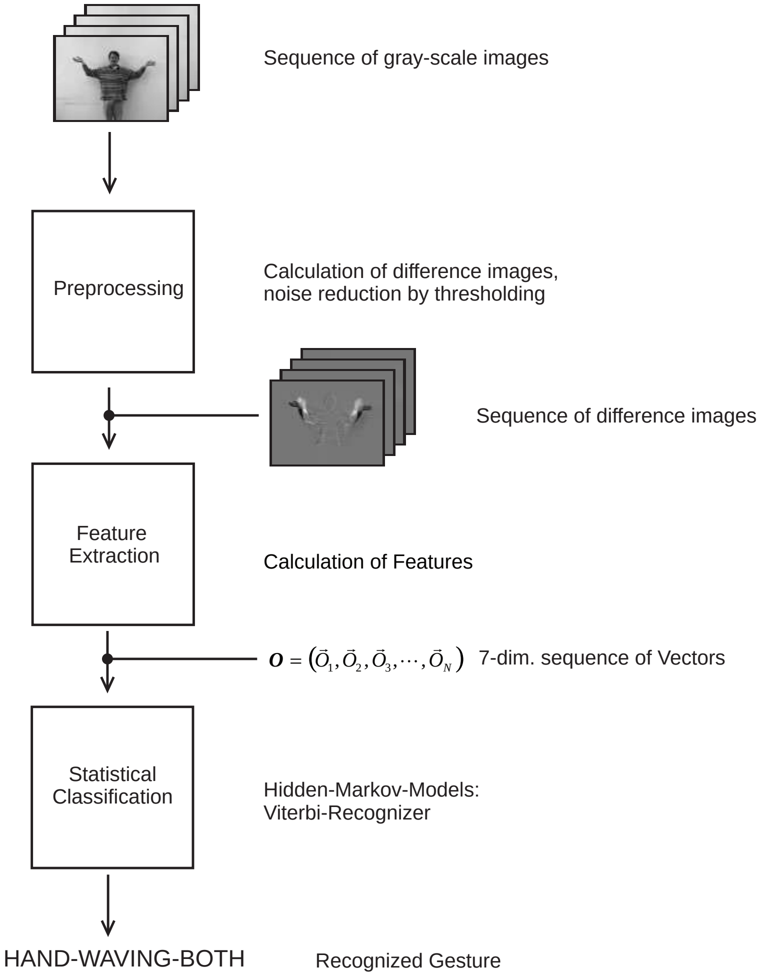

Sequence of gray-scale images
Preprocessing
Calculation of difference images, noise reduction by thresholding
Sequence of difference images
Feature Extraction
Calculation of Features
$\boldsymbol{O} = (\vec{O}_1, \vec{O}_2, \vec{O}_3, \cdots, \vec{O}_N)$ 7-dim. sequence of Vectors
Statistical Classification
Hidden-Markov-Models: Viterbi-Recognizer
HAND-WAVING-BOTH
Recognized Gesture

Fig. 2. Structure of the recognition system

proved to be more suitable for person-independent gesture recognition than the original image sequence. The difference image sequence is calculated by subtracting the pixel value at the same position (x, y) of adjacent frames of the original image sequence.

$$D'(x, y, t) = B(x, y, t) - B(x, y, t - 1) \tag{1}$$

This operation is shown in Figure 3, with the two adjacent original frames and the resulting difference image. The difference image contains positive and negative pixel values. Positive pixel values are black, negative values are white, and zero values are gray. Further can be seen in the difference image, that the background and static parts of the body are eliminated.

An easy way to reduce noise in the difference image is to apply a threshold operation to the difference image. Every pixel with an absolute value smaller than the threshold is set to zero.

$$D(x,y,t) = \begin{cases} 0 & : \quad |D'(x,y,t)| < S \\ D'(x,y,t) & : \quad |D'(x,y,t)| \geqq S \end{cases} \tag{2}$$

The size of the gray values $D(x,y)$ in this frame indicates the intensity of the motion for each spatial position (x,y) of the image.

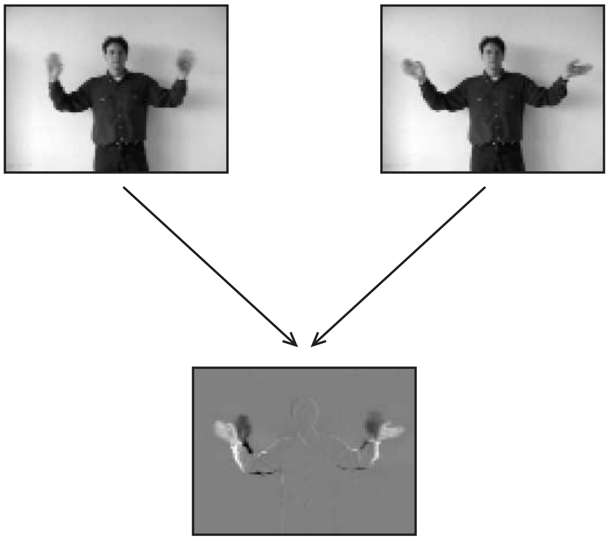


Fig. 3. Two original images and the resulting difference image

3.2 Feature Extraction

If one imagines the pixel values of the difference image as a "mountain area" of elevation $D(x,y)$ at point (x,y) (Figure 4), then this mountain area can be approximately considered as a distribution of the movement over the image space in x- and y-direction. Each mountain area will be characteristic for a specific motion, which may characterize a certain gesture. If it is possible to characterize this mountain area by certain features, these features should be a good representation for the current motion in the difference image. One possibility to express

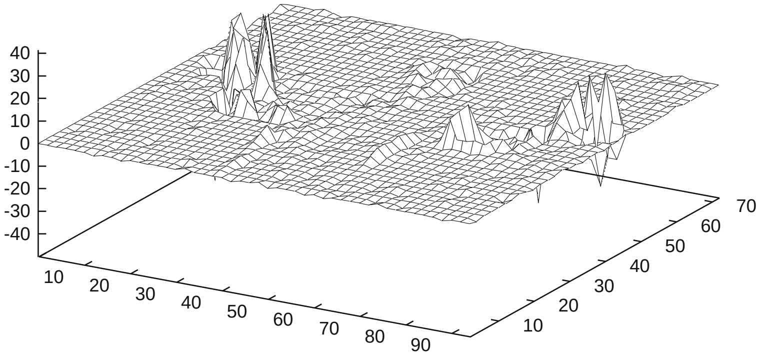


Fig. 4. Mountain area of the difference image of the gesture "Hand-Waving-Both"

this fact is the computation of the center of gravity $\mathbf{m}(t)^T = [m_x(t), m_y(t)]$ of the mountain area according to the formulas:

$$m_x(t) = \frac{\sum\limits_{x,y} x|D(x,y)|}{\sum\limits_{x,y} |D(x,y)|} \qquad\qquad m_y(t) = \frac{\sum\limits_{x,y} y|D(x,y)|}{\sum\limits_{x,y} |D(x,y)|} \tag{3}$$

The vector $\mathbf{m}(t)$ can then be also interpreted as "center of motion" of the image.

Another useful feature will be the mean absolute deviation of a pixel (x, y) from the center of motion $\boldsymbol{\sigma}(t)^T = [\sigma_x(t), \sigma_y(t)]$, with $\boldsymbol{\sigma}(t)$ defined as:

$$\sigma_x(t) = \frac{\sum\limits_{x,y} |D(x,y)(x-m_x(t))|}{\sum\limits_{x,y} |D(x,y)|} \qquad\qquad \sigma_y(t) = \frac{\sum\limits_{x,y} |D(x,y)(y-m_y(t))|}{\sum\limits_{x,y} |D(x,y)|} \tag{4}$$

This feature can be very helpful for distinguishing a gesture where large parts of the body is in motion (e.g. a "Kowtow") from a gesture concentrated more in a smaller area, where only a small body part moves (e.g. "nodding"). This feature can be also considered as "wideness of the movement". $\boldsymbol{\sigma}(t)$ is thus very similar to the variance, but is more robust against noise in the border area of the difference image.

Another important feature that should be included in this representation is the "intensity of motion" which is simply the average absolute height of the mountain area, expressed as

$$i(t) = \frac{\sum\limits_{x,y} |D(x,y)|}{\sum\limits_{x,y} 1} \tag{5}$$

where a large value of $i(t)$ represents a very intensive motion of large parts of the body, and a small value characterizes an almost stationary image.

If the "center of motion" is calculated separately for the positive and negative pixel values, the horizontal and vertical distance between the two centers can be used as additional features.

$$dm_x(t) = \frac{\sum\limits_{x,y|D(x,y)<0} xD(x,y)}{\sum\limits_{x,y|D(x,y)<0} D(x,y)} - \frac{\sum\limits_{x,y|D(x,y)>0} xD(x,y)}{\sum\limits_{x,y|D(x,y)>0} D(x,y)}$$

$$dm_y(t) = \frac{\sum\limits_{x,y|D(x,y)<0} yD(x,y)}{\sum\limits_{x,y|D(x,y)<0} D(x,y)} - \frac{\sum\limits_{x,y|D(x,y)>0} yD(x,y)}{\sum\limits_{x,y|D(x,y)>0} D(x,y)} \tag{6}$$

These features are an indication for the direction and the velocity of the moving object. They are useful to distinguish between the gestures nod-yes and nod-no. For both gestures the center, the wideness and the intensity of the motion are very similar. Only the direction is oriented horizontally for nodding-no and vertically for nodding-yes. It seems to be a good idea to separate the angle and the length of these features. But it should be considered, that it is almost impossible to model an angle by HMM, because of the $(\pi, -\pi)$ equivalence. Additionally, the angle has no meaning, if the distance of the two points is small, because the angle would change rapidly.

The use of $\mathbf{m}(t)$,$\boldsymbol{\sigma}(t)$ results in an interesting way for a visual representation of motion in the image as an ellipsis centered in $\mathbf{m}$ with the main axes $\boldsymbol{\sigma}$. The intensity of the motion $i(t)$ is represented by the brightness of the ellipsis. Figure 5 shows some gesture sequences and the overlaid ellipsis.

It should be noted, that this feature extraction method represents a remarkable reduction in complexity and dimension, by scaling down an image of $96 \times 72 = 6912$ grey values into a 7-dimensional vector

$$\mathbf{x_t}^T = (m_x, m_y, \sigma_x, \sigma_y, dm_x, dm_y, i), \tag{7}$$

while preserving the characteristics of the currently observed motion. This 7-dimensional motion vector is derived for each frame, resulting in a vector sequence $\mathbf{X} = \mathbf{x}_1, ..., \mathbf{x}_T$, where each vector carries important information about the current motion, and thus the entire sequence contains the information about the performed gesture.

3.3 Statistical Classification

The recognition of gestures in video sequences can be interpreted as a dynamic pattern recognition problem. Hidden Markov Models (HMMs) [7], well known from speech-recognition, offer superior pattern recognition capabilities for the dynamic case. Further advantages of using HMMs are, that segmentation and recognition takes place simultaneously, and that HMMs can be trained by a number of training samples similar to Neural Networks.

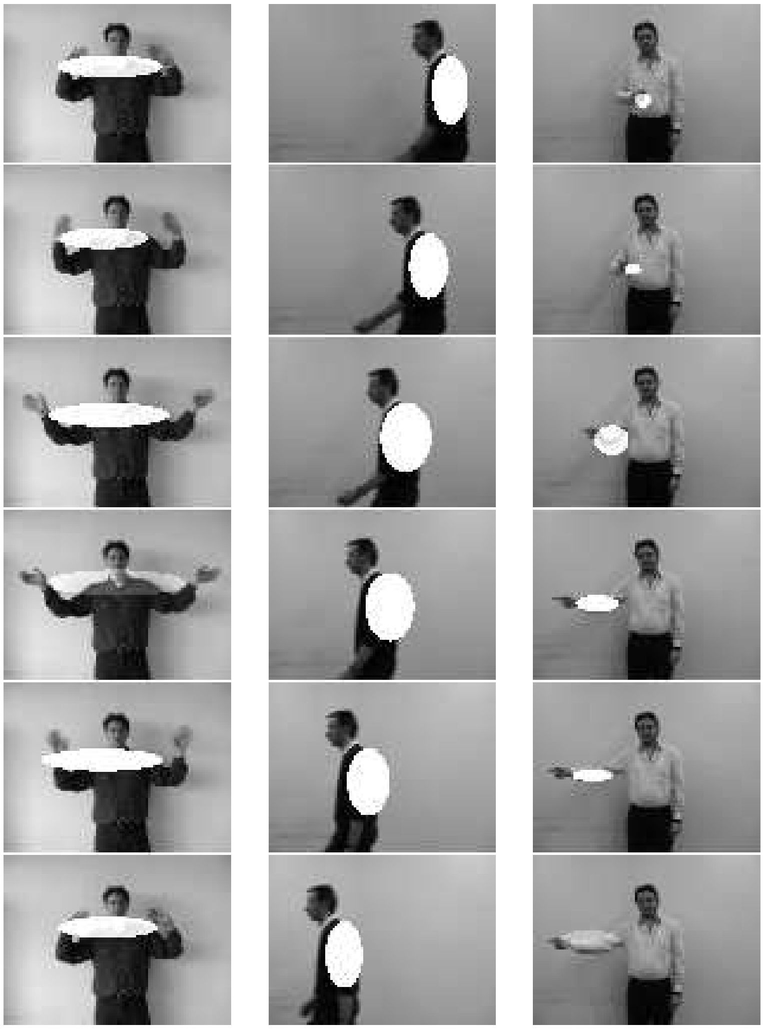


Fig. 5. Images of the gestures Hand-Waving-Both, Go-Right and Round-Clockwise

The system presented here uses 24 different HMMs for 24 different gestures. The HMM parameters were estimated with the feature sequences of the according gesture samples, applying the Forward-Backward algorithm. As shown in Figure 6, a HMM λ consists of a number of states (four states in Figure 6) with probability density functions (pdfs) and transition probabilities (characterized as arrows). Modeling a seven dimensional feature sequence requires a seven dimensional Gaussian pdf.

The recognition result is the HMM with the highest probability for the given observation sequence, while each frame of the observation sequence is aligned to a HMM-state. The resulting probability $P(\mathbf{X}|\lambda_l)$ for a HMM λ_l with a given observation sequence $\mathbf{X}$ is calculated with the Viterbi algorithm, which delivers the probability of the most likely state sequence for the HMM λ_l.

With respect to the different types of gestures, like periodical or linear gestures, different HMM topologies are used. Linear gestures (gestures 17-24), like "go right" for instance are modeled with a linear topology (Figure 7a), where each state has only self-transitions and transitions to the following state. Cyclic, or periodical gestures (gestures 1-16), like "round clockwise" or "pointing to top", which appear more than once in a sequence, are modeled with a cyclic topology, as shown in Figure 7b.

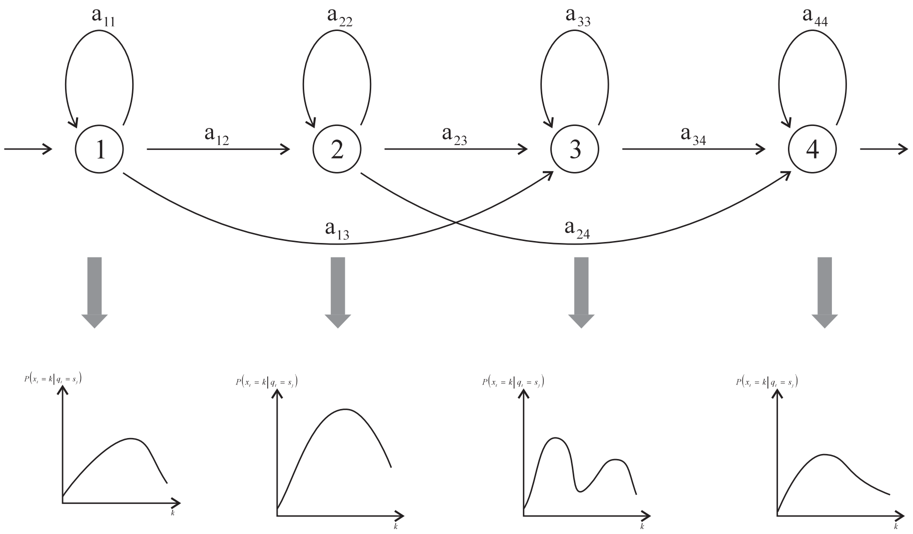


Fig. 6. Structure of a Hidden Markov Model

4 Results

The recognition rate for the task, presented in Chapter 2 is 92.9%. Figure 8 shows the recognition rates for the different gestures. Even the very difficult gestures like "drawing a letter" have recognition rates, that are only slightly lower than the average recognition rate. "Nodding yes and no" are two difficult gestures, too. Both contain only very little motion.

For applications of gesture recognition, like robot control for instance, it is important to reject sequences, which do not belong to the "gesture vocabulary". To solve this problem we recorded a few additional movements like folding arms and scratch the head but also movements of two people, as well as non movement (silence). Several garbage models were trained by these garbage gestures. A test of this rejection method showed, that the recognition results for the 24 gestures decreased only a little and most of the movements outside the gesture set were rejected.

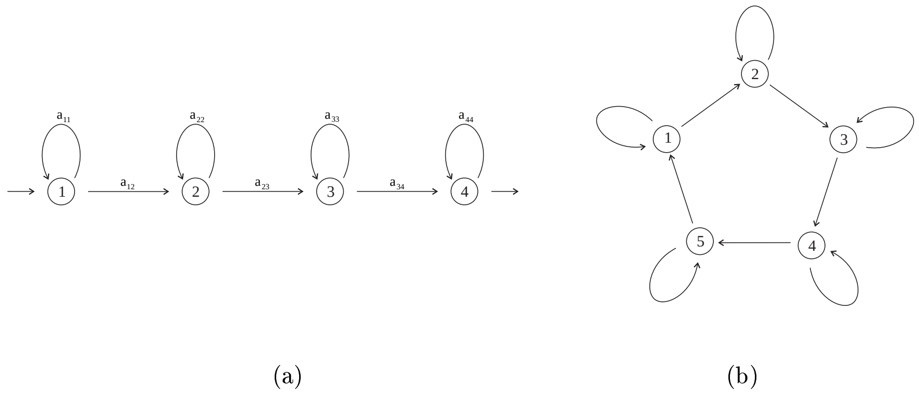


Fig. 7. (a) Linear and (b) cyclic model

5 Conclusions

A new gesture recognition method based on Hidden Markov Models was presented. We introduced a feature extraction method for global motion features, that reduces the amount of data significantly, while the motion information of the moving object in the sequence is preserved. Because of the data reduction capability it is possible to recognize the gestures in real-time with high accuracy. The feature extraction from the difference image enables person and background independent recognition. Furthermore the recognition is robust against illumination changes. The system can be easily extended for the recognition of additional gestures, because of the automatic learning capabilities of the Hidden Markov Model approach. Many other applications in the field of video sequence processing, like video indexing [8] for instance, can be found for this flexible approach.

Further improvements of the system are expected for the extension to position invariant recognition. For real-world applications it is important to recognize the gestures continuously. This means the simultaneous recording and recognition of a video sequence of infinite length, while the currently recognized gesture is displayed immediately. Activities in this area are already in progress and results will be presented in the near future.

References

1. Claudette Cédras and Mubarak Shah. Motion-Based Recognition: A Survey. *Image and Vision Computing*, 13(2):129–155, 1995.

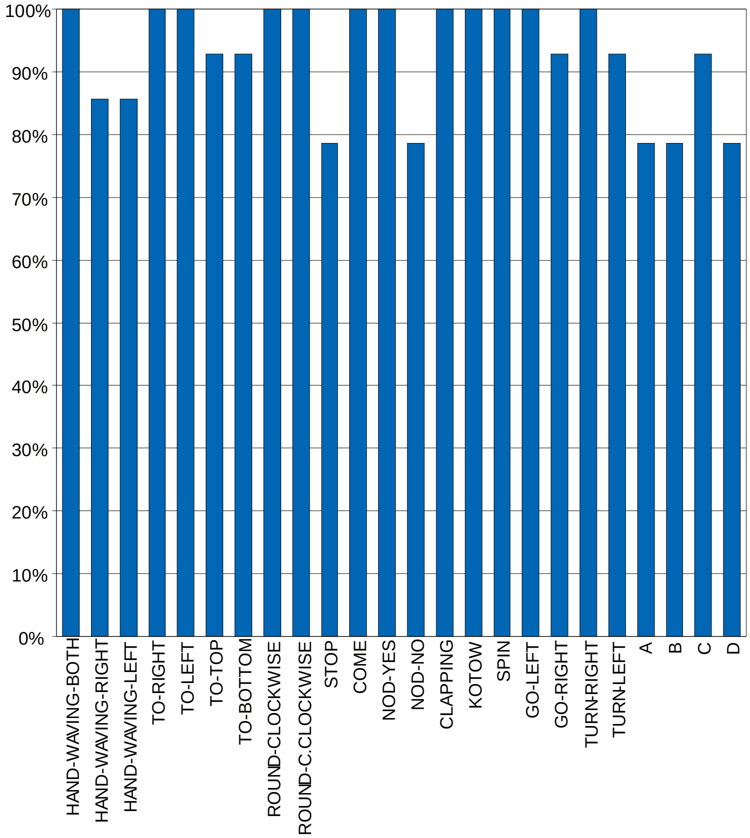


Fig. 8. Recognition rates for the different gestures

2. Aaron Bobick and James Davis. An Apperance-Based Representation of Action. In *Proc. IEEE Int. Conf. on Pattern Recognition (ICPR-96)*, pages 307–312, Vienna, Austria, August 1996.
3. Thad Starner and Alex Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models. In *International Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, 1995.
4. Polly K. Pook. Teleassistance: Using deictic gestures to control robot action. Technical Report TR594, University of Rochester, Computer Science Department, September 1995.
5. Mike Schuster and Gerhard Rigoll. Fast Online Video Image Sequence Recognition with Statistical Methods. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3450–3453, Atlanta, May 1996.
6. Gerhard Rigoll and Andreas Kosmala. New improved Feature Extraction Methods for Real-Time High Performance Image Sequence Recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2901–2904,

Munich, April 1997.

7. Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of the IEEE*, 77(2):257–285, 1989.
8. Stefan Eickeler, Andreas Kosmala, and Gerhard Rigoll. A New Approach to Content-Based Video Indexing Using Hidden Markov Models. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 149–154, Louvain-la-Neuve, Belgium, June 1997.