Modeling Image Variability in Appearance-Based Gesture Recognition

Philippe Dreuw, Thomas Deselaers, Daniel Keysers, and Hermann Ney

Lehrstuhl für Informatik 6 – Computer Science Department,
RWTH Aachen University – D-52056 Aachen, Germany
{dreuw, deselaers, keysers, ney}@informatik.rwth-aachen.de

Abstract. We introduce the use of appearance-based features in hidden Markov model emission probabilities to recognize dynamic gestures. Tangent distance and the image distortion model are used to directly model image variability in videos. No explicit hand models and no segmentation of the hand is necessary. Different appearance-based features are investigated and the invariant distance measures are systematically evaluated. The approach is evaluated for three tasks of strongly varying difficulty and and performs favorably well. We obtain promising first results on a novel database of the German finger-spelling alphabet.

1 Introduction

In sign language, the gestures are part of a visual language and well defined. The gestures are used to communicate in the form of finger-spelling, as complete words, or as non-manual features. Many disciplines must be combined to achieve a reliable recognition system, as one has to deal with e.g. capturing problems like varying lighting conditions, skin colored clothes, or tracking of multiple objects.

Work in the field of vision-based gesture recognition usually first segments parts of the input images, for example the hand, and then uses features calculated from this segmented input like shape or motion [3]. Problems with this approach are tracking, occlusion, lighting, or clothing constraints.

Results in the field of object recognition in images suggest that this intermediate segmentation step may not be necessary. The question addressed in our research is if appearance based features are competitive for gesture recognition and if we can use similar models of image variability as in object recognition. The experiments presented in this work will show that the answer to this question is positive. We also want to know which features are suitable and what are the appropriate choices for the hidden Markov model (HMM) parameters.

The main focus of this work is set on using appearance-based features with no need for complex feature extraction. We integrated distance measures known from image and optical character recognition (e.g. being invariant against affine transformations) into the hidden Markov model classifiers to model image variability.

2 Related Work

One of the first "working" real-time sign language recognition systems was developed in [17]. The authors' HMM-based system works without explicitly modeling the fingers and recognizes American sign language on a sentence level. The tracking module can be used with or without colored gloves, where the resultant shape, orientation, and trajectory information are taken as input features to an HMM for recognition. With a 40 word lexicon, an error rate of 8% for the skin color tracking case is achieved.

A person-independent real-time system for gesture recognition is presented in [16]. The system uses global motion features extracted from each difference image of the image sequence, and HMMs as a statistical classifier. These HMMs are trained on a database of 24 isolated gestures, performed by 14 different people. An error rate of 7.1% is achieved, but the system can only distinguish between gestures that can be characterized by their movement.

In [2], a view-based approach to the representation and recognition of human movement using temporal templates is presented. The authors develop a recognition method by matching temporal templates against stored instances of views of known actions.

The gesture recognition system presented in [11] can recognize a vocabulary of 46 single-hand gestures of the American sign language finger spelling alphabet and digits in real time. Each video frame is processed independently, and dynamic gestures are replaced by static ones. The system was trained and tested using data of one person and thus is highly person dependent.

A two-stage classification procedure is presented in [3] where an initial classification stage extracts a high-level description of hand shape and motion. A second stage of classification is then used to model the temporal transitions of individual signs using a classifier bank of Markov chains combined with Independent Component Analysis.

In [4] a classification system using global features is presented. In contrast to the work presented here, the training data are manually segmented and only the relevant part of the video, i.e. exactly the frames where a sign is gestured, is taken into account.

Most of the systems presented here assume a constant environment for their systems, e.g. persons wearing non-skin-colored clothes with long sleeves and a fixed camera position under constant lighting conditions. The presented systems are often very person-dependent and the used gestures exhibit large differences to be easily recognizable.

3 Appearance-Based Features for Gesture Recognition

In an appearance-based approach the image itself and simple transformations (filtering, sub-sampling, ...) of the image are usually used as features. In this paper, we will denote an original image X in a sequence at time $t = 1, ..., T$ by X_t, and the pixel value at the position (x, y) by $X_t(x, y)$. Any derived image will be denoted by $\widetilde{X}$.

(a) (b) (c) (d) (e) (f) (g)

Fig. 1. Original appearance based features and spatial derivatives (from left to right): original, filtered with horizontal Sobel filter, filtered with vertical Sobel filter and filtered with magnitude Sobel filter. Difference images: absolute first-order time derivative, first-order time derivative and second-order time derivative.

Original Images. When working, for example, with gray valued images (e.g. infrared images like in Fig. 1(a)), a (thresholded) original image can be used as a feature. Using original image sequences or their spatial derivatives as a feature without any thresholding or tracking can already lead to very good results.

Difference Images. Calculating difference images is one of the simplest methods of detecting motion in an image sequence. Motion is a very important appearance-based feature in image sequences, which captures the relation between local properties and time variation. This method is fast, and the optical flow in the motion field can be used in further processing steps and applications.

The first-order time derivative difference image $\widetilde{X}_t$ (see Fig. 1(f)), corresponding to the original image X_t, is calculated as follows:

$$\widetilde{X}_t(x,y) = X_{t+1}(x,y) - X_{t-1}(x,y) \tag{1}$$

The second-order time derivative difference image $\widetilde{\widetilde{X}}_t$ (see Fig. 1(g)), corresponding to the original image X_t, is calculated as follows:

$$\widetilde{\widetilde{X}}_t(x,y) = X_{t-1}(x,y) - 2 \cdot X_t(x,y) + X_{t+1}(x,y) \tag{2}$$

Motion History. The notions motion-energy-image (MEI) and motion-history-image (MHI) were introduced in [2]. The basic idea is to construct an image that can be matched against stored representations of known movements. This image is used as a temporal template.

To represent *how* (as opposed to *where*) motion in the image is moving, an MHI is formed. In an MHI H_t, the pixel intensity is a function of the temporal history of motion at that point, and a simple replacement and decay operator τ is used (with $1 \leq \tau \leq N$ for a sequence of length N):

$$H_t(x,y) = \begin{cases} \tau & \text{if } |\widetilde{X}_t(x,y)| > T_0 \\ \max(0, H_{t-1}(x,y) - 1) & \text{otherwise} \end{cases} \tag{3}$$

The result is a scalar-valued image where more recently moving pixels are brighter. Note that the MEI can be generated by thresholding the MHI above zero. Fig. 2 shows a key frame with its corresponding MHI and MEI.

Fig. 2. Motion energy and history image examples on the DUISBURG-Gesture database: the original key frame at time $t = 47$ of the gesture "Round-Clockwise" with the corresponding motion-history-image and motion-energy-image.

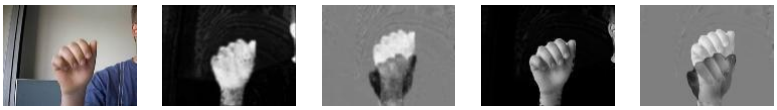


Fig. 3. Skin color image features: original, skin probability, 1^{st} time derivative of skin probability, original thresholded by skin probability and 1^{st} time derivative of original thresholded by skin probability.

Skin Color Images. The skin color model used in this paper is based on the Compaq Cambridge Research Lab image-database presented in [8]. Skin probability images denoted as S were created according to their skin probability maps, and read from the given skin- and non-skin color models which were estimated from a larger collection of pictures.

Therefore, one can also segment the original image X by its own skin color probability where T_p is a suitable skin color probability threshold. Applying a Gaussian filter on the skin color probability map before thresholding the original image can improve the segmentation as gaps in contiguous skin regions are reduced. Instead of a fixed threshold, the thresholding can be improved once more by using a sigmoid function:

$$X(x, y) = \frac{1}{1 + \exp(-\alpha \cdot (S(x, y) - T_p))} \tag{4}$$

These Gaussian and sigmoid smoothing functions to segment skin regions are not necessarily the optimal methods and many alternative algorithms have been suggested [15,19]. Fig. 3 shows some examples of possible features derived from skin color probability maps.

4 Hidden Markov Models

The ability of Hidden Markov models to compensate time and amplitude variations has been proven for speech recognition [7], gesture recognition, [13], sign language recognition [17,18] and human action recognition [5,12]. We focus especially on distance measures being invariant against slight affine transformations or distortions.

Problems that have an inherent temporality may have states at time t that are influenced directly by a state at time $t - 1$. The idea of a HMM is to represent a

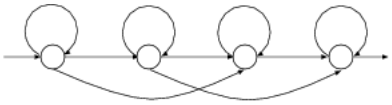


Fig. 4. (0,1,2)-Standard model where the discrete states s are represented by nodes and the transition probabilities by links.

signal by a state of a stochastic finite state machine. A more detailed description can be found in [7].

To classify an observation sequence X_1^T, we use the Bayesian decision rule:

$$\begin{aligned} X_1^T \longrightarrow r(X_1^T) &= \underset{k}{\operatorname{argmax}} \left\{ p(k|X_1^T) \right\} = \underset{k}{\operatorname{argmax}} \left\{ p(k) \cdot p(X_1^T|k) \right\} \\ &\overset{\text{model}}{\cong} \underset{k}{\operatorname{argmax}} \left\{ p(k) \cdot \max_{s_1^T} \left\{ \prod_{t=1}^{T} \underbrace{p(s_t|s_{t-1},k)}_{\text{transition}} \cdot \underbrace{p(X_t|s_t,k)}_{\text{emission}} \right\} \right\} \end{aligned} \quad (5)$$

where X_1^T is a sequence with images $X_1, ..., X_t, ..., X_T$. Here, $p(k)$ is the a priori probability of class k, $p(X_1^T|k)$ is the class conditional probability for the observation X_1^T given class k and $r(X_1^T)$ is the decision of the classifier.

We only use linear models in this work, e.g. the 0-1 model which allows loop and forward transitions, and the 0-1-2 model which additionally allows skip transitions. Fig. 4 shows a 0-1-2-standard HMM topology.

It is necessary to choose models for the respective distributions and estimate their parameters using training data. The emission probabilities are modeled using Gaussian mixture densities in the experiments presented later. We assume Σ to be diagonal:

$$p(X|s) = \sum_{i=1}^{l_k} \mathcal{N}(X|\mu_i, \Sigma) \quad (6)$$

In each state s of an HMM, a distance is calculated. We assume pooled variances over all classes and states, i.e. we use $\sigma_{sdk} = \sigma_d$. The negative logarithm of $p(X|s)$ can be interpreted as a distance $d(p(X|s))$ and is used as emission score:

$$-\log(p(X|s)) = \frac{1}{2} \left(\sum_{d=1}^{D} \left(\underbrace{\left(\frac{X_d - \mu_{sd}}{\sigma_d} \right)^2}_{\text{distance}} + \underbrace{\log(2\pi\sigma_d^2)}_{\text{normalization factor}} \right) \right) \quad (7)$$

When working with image sequences, we calculate a distance between two images, e.g. we compare the current observation image X_t (or any transformed image $\widetilde{X}_t$) with the mean image μ_s at this state. Simply comparing the pixel values is quite often used in object recognition but different methods have been proposed to do this.

One of the main topics in this paper is the use of different distance measures inside the HMM's emission probabilities to model image variability. As in

character or image recognition, we want to analyze whether transformation independent distance measures can improve the recognition performance. Usually normalized distance measures are used:

$$d(X, \mu_s) = \sum_{d=1}^{D} \left(\frac{X_d - \mu_{sd}}{\sigma_d} \right)^2 \tag{8}$$

The Euclidean distance has been successfully used e.g. in optical character and object recognition and has been extended by different methods. This distance measure will be replaced by the tangent distance or the image distortion model.

Tangent Distance. Because the Euclidean distance does not account for affine transformations such as scaling, translation and rotation, the tangent distance (TD), as described in [10], is one approach to incorporate invariance with respect to certain transformations into a classification system. Here, invariant means that image transformations that do not change the class of the image should not have a large impact on the distance between the images.

Let $X \in \mathbb{R}^D$ be a pattern and $T(X, \alpha)$ denote a transformation of X that depends on a parameter L-tuple $\alpha \in \mathbb{R}^L$. We assume that T does not change class membership (for small α). The manifold of all transformed patterns $\mathcal{M}_X = \{T(X, \alpha) : \alpha \in \mathbb{R}^L\} \subset \mathbb{R}^D$ now offers new possibilities for distance calculations. The distance between two patterns X and μ can be defined as the minimum distance between the two manifolds $\mathcal{M}_X$ and $\mathcal{M}_\mu$, which is truly invariant with respect to the regarded transformations.

The distance calculation between manifolds is a hard non-linear optimization problem in general. These manifolds can be approximated by a tangent subspace $\widehat{\mathcal{M}}$ which is spanned by a set of tangent vectors X^l which are the partial derivatives of the transformation T with respect to the parameters α_l. Thus, the transformation $T(X, \alpha)$ can be approximated using a Taylor expansion around $\alpha = 0$.

$$t(X, \alpha) = X + \sum_{l=1}^{L} \alpha_l X_l + \sum_{l=1}^{L} \mathcal{O}(\alpha_l^2) \tag{9}$$

The set of points consisting of all linear combinations of the tangent vectors X^l in the point X forms the tangent subspace $\widehat{\mathcal{M}_X}$ as a first-order approximation.

Using the linear approximation $\widehat{\mathcal{M}_X}$ allows us to calculate the distances as a solution of a least squares problem or projections into subspaces. Both are computationally inexpensive operations. The approximation is valid for small values of α, which nevertheless is sufficient in many applications. Patterns that all lie in the same subspace can be therefore represented by one prototype and the corresponding tangent vectors. The TD between the original image and any of the transformations is therefore zero, while the Euclidean distance is significantly greater than zero. Using the squared Euclidean norm, the TD is defined as:

$$d_{2\mathrm{S}}(X, \mu_s) = \min_{\alpha, \beta \in \mathbb{R}^L} \left\{ ||(X + \sum_{l=1}^{L} \alpha_l \mu_{sl}) - (\mu + \sum_{l=1}^{L} \beta_l \mu_{sl})||^2 \right\} \tag{10}$$

This distance measure is also known as a two-sided tangent distance (TD2S). To reduce the effort for determining $d_{2\mathrm{S}}(X, \mu)$, it may be convenient to restrict the

(a) (b)

Fig. 5. (a) The tangent vectors corresponding to the six affine transformations horizontal shift, vertical shift, first and second hyperbolic transformation, scaling and rotation of a mean image μ (from the RWTH-Gesture database) used to create the transformed mean image. (b) An observation X, a tangent transformed mean image and the original mean image μ, achieved by minimizing the tangent distance TD1S.

tangent subspaces to the derivatives of the reference (or the observation), which results in a one-sided tangent distance (TD1S) (see Fig. 5).

Image Distortion Model. The image distortion model [9] is a method which allows for small local deformations of an image. Each pixel is aligned to the pixel with the smallest squared distance from its neighborhood. These squared distances are summed up for the complete image to get the global distance. To compare an observation image X_t with a mean image μ_{s_t}, $d(X_t, \mu_{s_t})$ is calculated as follows:

$$d_{\text{idm}}(X, \mu_s) = \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} \min_{x'=x-w}^{x+w} \min_{y'=y-w}^{y+w} d'(X(x,y), \mu_s(x', y')) \tag{11}$$

Here, w is the warp range, i.e. the radius of the neighborhood in which a pixel may be chosen for alignment, and d' is a pixel distance comparing the image pixels $X_t(x,y)$ and $\mu_{s_t}(x', y')$ for example the Euclidean distance. This method can be improved by enhancing the pixel distance d' to compare sub images of size $(2v+1) \times (2v+1)$ instead of single pixels only:

$$d'(X(x,y), \mu_s(x', y')) = \sum_{i=-v}^{v} \sum_{j=-v}^{v} (X(x+i, y+j) - \mu_s(x'+i, y'+j))^2 \tag{12}$$

Further improvement is achieved by using spatial derivatives instead of the pixel values directly. Intuitively, the use of derivatives makes the image distortion model align edges to edges and homogeneous areas to homogeneous areas. Fig. 6 shows some examples of distorting mean images with respect to observations so that their pixel distance is minimal.

5 Databases

In this section we present the databases used to benchmark our system.

Fig. 6. IDM distortion example on the RWTH-Gesture database: observation X, distorted mean image with the smallest distance $d'(X, \mu_s)$, original mean image μ_s, vertical and horizontal Sobel images used for distortion.

Fig. 7. Some examples of the LTI-Gesture database.

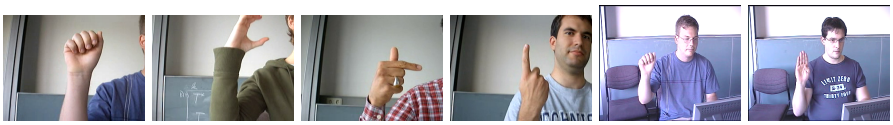
LTI-Gesture Database. The LTI-Gesture database was created at the Chair of Technical Computer Science of the RWTH Aachen University [1]. It contains 14 dynamic gestures, 140 training and 140 testing sequences. An error rate of 4.3% was achieved on this database in [14]. HMMs are required for recognition as some gestures can only be distinguished using motion. In particular, the gestures 'five', 'stop', and 'pause' have the same hand shape but differ in the movement of the hand.

DUISBURG-Gesture Database. For the training and the testing of the system presented in [16] video sequences of 24 different dynamic gestures were recorded. The resolution of the video sequences was 96 x 72 gray-scale pixel and 16 frames per second. Fig. 8 shows some examples of the different gestures. The database consists of 336 image sequences that contain gestures of 12 different persons. With a leaving-one-person-out classification an error rate of 7.1% was achieved.

RWTH-Gesture Database. We recorded a database of finger spelling gestures of German Sign Language. Our database is freely available on our website[1]. The database contains 35 gestures with video sequences showing the signs 'A' to 'Z', 'SCH', the German umlauts 'Ä', 'Ö', 'Ü', and the numbers '1' to '5'. HMMs are necessary for recognition as five of the gestures contain inherent motion ('J', 'Z', 'Ä', 'Ö', and 'Ü'). The database consists of disjunct sets of 700 training sequences and 700 test sequences. In total 20 signers occur in the sequences.

[1] `http://www-i6.informatik.rwth-aachen.de/~dreuw/database.html`

Fig. 8. Some examples of the DUISBURG-Gesture database.



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

Fig. 9. Some examples of the RWTH-Gesture database showing different gestures of letters: (a)-(d) recorded with the webcam with (a) "A", (b) "C", (c) "T", and (d) "1". (e)-(f) are recorded with the camcorder with (e) "A" and (f) "B".

The recording was done under non-uniform daylight lighting conditions, the camera viewpoints are not constant, and the persons had no restrictions on the clothing while gesturing. Each person signed each gesture twice on two different days. The gestures were recorded by two different cameras, one webcam (resolution 320×240) and one camcorder (resolution 352×288) at 25 frames per sconed, from different points of view. Fig. 9 shows some examples of the different gestures. More information about the database is available on our website.

6 Results

We made some basic experiments in [6] on the LTI-Gesture database to determine the parameters for the HMM, necessary to recognize the gestures which contain inherent motion. We summarize the results here briefly: We found that using Gaussian mixture densities, a 0-1-2 model, and pooling over the variances achieved the best results. Pruning of hypotheses can improve the run-time by a factor of 4. We also made experiments about the relative weight between transition and emission score. The emission score weight is the exponent of the emission probability in Eq. 5. One can conclude from the results in Fig. 10 that, in the task of recognizing image sequences, a high emission score weight is very important.

In [14], an error rate of 4.3% was achieved for the LTI-Gesture database using shape and motion features in combination with forearm segmentation. Using the centroid features as presented in [16], we have only achieved an error rate of 14.2%, and we can conclude that these features should only be used to describe motion patterns instead of more complex hand shapes. Using original image features on the LTI-Gesture database, we have improved the error rate of 5.7% to 1.4% in combination with the tangent distance [6]. Using the IDM we have also achieved an error rate of 1.4% (see Tab. 1).

We achieved an error rate of as high as 61.7% using original image features on the DUISBURG-Gesture database, which was expected due to the full-body gestures, i.e. the different clothing had a high impact on the error rate. With the absolute 1^{st} time derivative image feature, we achieved an error rate of 14.2% which has also been improved with tangent distance to the competitive error rate of 13.2%. Furthermore the performance of the MHI images has been improved

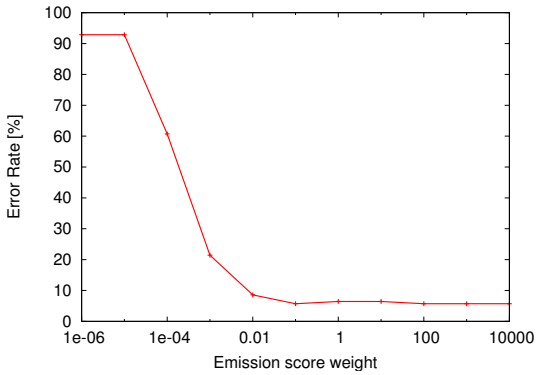


Fig. 10. Error rates[%] with 32x32 original features on LTI-Gesture database and estimated transition probabilities for 0-1-2 model against emission score weight showing that a high emission score weight yields the best results

Table 1. Error rates [%] on the LTI-Gesture database.

Features	Euclidean	Tangent	IDM
COG	14.2	–	–
original	5.7	1.4	1.4
vertical Sobel	5.0	2.8	1.4
magnitude Sobel	7.1	1.4	1.4
motion-history	5.7	3.5	6.4

with the new distance measures (see Tab. 2) both for the HMM approach and the template based approach as described in [2].

On the RWTH-Gesture database, we used only the webcam images to test our system. Since the camera position is not constant, the signing persons do not wear the same clothing, and the lighting conditions are changing, we decided to make a first test with full size skin thresholded original image features down-scaled to 32x32. With this feature we achieved an error rate of 87.1%. Using the 1^{st} time derivative of original images thresholded by their skin probability, we have achieved an error rate of 72.1%.

It is obvious that this database contains gestures of very high complexity and recognition is also complicated by the very high inter-class similarity of many gestures. Therefore, we need additional methods for feature extraction or other

Table 2. Error Rates [%] on the DUISBURG-Gesture database.

Features	Euclidean	Tangent	IDM
absolute 1^{st} time der.	14.2	13.2	-
motion-history (HMM)	18.7	16.9	-
motion-history (Template)	20.7	19.0	17.5

Table 3. Error Rates [%] on the RWTH-Gesture database.

Feature	Euclidean	Tangent
original thresholded by skin color prob. (i.e. image intensity) (*)	87.1	-
+ camshift tracking (no segmentation)	44.0	35.7
1^{st} time derivative of (*) (i.e. spatial differences)	72.1	-
+ camshift tracking (no segmentation)	46.2	44.1

distance measures. Using a camshift tracker to extract more position independent features (note that we do not try to segment the hand), we have improved the error rate from 87.1% to 44.0% using the original images thresholded by their skin probability. With the 1^{st} time derivative image feature of original images thresholded by their skin probability in combination with tracking, the error rate has been improved from 72.1% to 46.2%.

Using a two-sided tangent distance we have improved the error rate to the currently best result of 35.7%, which shows the advantage of using distance measures that are invariant against small affine transformations and the possibility of recognizing gestures by appearance-based features. We also have improved the error rate when using the 1^{st} time derivative image feature of original images thresholded by their skin probability with two-sided tangent distance from 46.2% to 44.1%. Fig. 3 shows the achieved results on this database up to now.

7 Conclusion

We presented an approach to the recognition of dynamic gestures that uses several appearance-based features with distance functions that are invariant with respect to certain transformation in an HMM-based recognition framework. The approach is evaluated on three different tasks and performs favorably well.

The databases tasks addressed are of strongly varying difficulty where the simplest task of one-handed gesture recognition in a controlled environment can be considered solved, the results for the medium-hard task are competitive to results that were obtained with a method optimized with this respect to this task. For the recognition of more complex gestures in the finger spelling domain, we showed that the approach is suitable and that further improvements can be expected in the near future.

The best achieved error rate on the RWTH-Gesture database so far is 35.7% which shows the high complexity of this database. Nevertheless, this result is promising because only a simple webcam without any restriction for the signer was used and some signs are visually very similar, as for example the signs for 'M', 'N', 'A', and 'S' (cp. Fig. 11).

Furthermore, it has been shown that the tangent distance and the image distortion model can suitably be integrated into an HMM-based recognition framework and that the advantages of these invariant distance functions that have in the past been successfully exploited in the domain of still images can directly be transfered to the recognition of videos.

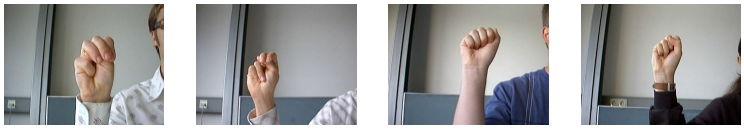


Fig. 11. Some examples of visually very similar signs “M”, “N”, “A”, and “S” of the RWTH-Gesture database.

Some questions still remain unanswered, e.g. not all distance measures were completely analyzed in combination with tracking on the RWTH-Gesture database and the combination of different features was not yet completely performed.

References

1. S. Akyol, U. Canzler, K. Bengler, and W. Hahn. Gesture Control for Use in Automobiles. In *IAPR WMVA 2000*, Tokyo, Japan, pages 349–352, Nov 2000. 8
2. A. F. Bobick and J. W. Davis. The Recognition of Human Movement Using Temporal Templates. *IEEE PAMI 2001*, 23(3):257–267, Mar 2001. 2, 3, 10
3. R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A Linguistic Feature Vector for the Visual Interpretation of Sign Language. In J. M. Tomas Pajdla, editor, *ECCV 2004*, volume 1, Prague, Czech Republic, pages 391–401, May 2004. 1, 2
4. S.-F. Wong and R. Cipolla Real-time Interpretation of Hand Motions using a Sparse Bayesian Classifier on Motion Gradient Orientation Images. In *BMVC 2005*, vol. 1, Oxford, UK, pages 379–388, Sept. 2005. 2
5. M. Brand, N. Oliver, and A. Pentland. Coupled hidden Markov models for complex action recognition. In *IEEE CVPR 1997*, IEEE Computer Society, Washington, DC, USA, pages 994–, Jun 1997. 4
6. P. Dreuw. Appearance-Based Gesture Recognition. Diploma thesis, RWTH Aachen University, Aachen, Germany, Jan 2005. 9
7. F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, Massachusetts, Jan 1998. 4, 5
8. M. Jones and J. Rehg. Statistical Color Models with Application to Skin Color Detection. Technical Report CRL 98/11, Compaq Cambridge Research Lab, 1998. 4
9. D. Keysers, J. Dahmen, H. Ney, B. Wein, and T. Lehmann. Statistical Framework for Model-based Image Retrieval in Medical Applications. *Journal of Electronic Imaging*, 12(1):59–68, Jan 2003. 7
10. D. Keysers, W. Macherey, H. Ney, and J. Dahmen. Adaptation in Statistical Pattern Recognition using Tangent Vectors. *IEEE PAMI 2004*, 26(2):269–274, Feb 2004. 6
11. R. Lockton and A. W. Fitzgibbon. Real-time gesture recognition using deterministic boosting. In *BMVC 2002*, Cardiff, UK, pages 817–826, Sep 2002. 2
12. N. Nguyen, H. Bui, S. Venkatesh, and G. West. Recognising and monitoring high-level behaviours in complex spatial environments. In *IEEE CVPR 2003*, volume 2, Madison, Wisconsin, pages 620–625, Jun 2003. 4
13. V. Pavlovic, R. Sharma, and T. S. Huang. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. *IEEE PAMI 1997*, 19(7):677–695, Jul 1997. 4
14. A. Pelkmann. Entwicklung eines Klassifikators zur videobasierten Erkennung von Gesten. Diploma thesis, RWTH Aachen University, Aachen, Germany, Feb 1999. 8, 9

15. Y. Raja, S. J. McKenna, and S. Gong. Tracking and Segmenting People in Varying Lighting Conditions using Colour. In *3rd IEEE FGR 1998*, Nara, Japan, pages 228–233, Apr 1998. 4
16. G. Rigoll, A. Kosmala, and S. Eickeler. High Performance Real-Time Gesture Recognition using Hidden Markov Models. In *International Gesture Workshop*, volume 1371, Springer-Verlag, Bielefeld, Germany, pages 69–80, Sep 1998. 2, 8, 9
17. T. Starner, J. Weaver, and A. Pentland. Real-time American sign-language recognition using desk and wearable computer based video. *IEEE PAMI*, 20(12):1371–1375, Dec 1998. 2, 4
18. C. Vogler and D. Metaxas. A Framework for Recognizing the Simultaneous Aspects of American Sign Language. *CVIU 2001*, 81(3):358–384, Mar 2001. 4
19. X. Zhu, J. Yang, and A. Waibel. Segmenting Hands of Arbitrary Color. In *AFGR 2000*, Grenoble, France, pages 446–453, Mar 2000. 4