The Significance of Facial Features for Automatic Sign Language Recognition

Ulrich von Agris, Moritz Knorr, and Karl-Friedrich Kraiss
Institute of Man-Machine Interaction, RWTH Aachen University, Germany
{vonagris,knorr,kraiss}@mmi.rwth-aachen.de

Abstract

Although facial features are considered to be essential for humans to understand sign language, no prior research work has yet examined their significance for automatic sign language recognition or presented some evaluation results. This paper describes a vision-based recognition system that employs both manual and facial features, extracted from the same input image. For facial feature extraction an active appearance model is applied to identify areas of interest such as the eyes and mouth region. Afterwards a numerical description of facial expression and lip outline is computed. An extensive evaluation was performed on a new sign language corpus, which contains continuous articulations of 25 native signers. The obtained results proved the importance of integrating facial expressions into the classification process. The recognition rates for isolated and continuous signing increased in signer-dependent as well as in signer-independent operation mode. Interestingly, roughly two of ten signs were recognized just from the facial features.

1. Introduction

Sign language is the natural language of deaf and hard of hearing people used for everyday communication among themselves. Although different in form, it serves the same functions as a spoken language. Spread all over the world, it is not a universal language. Regionally different languages have been evolved such as American Sign Language (ASL) and German Sign Language (DGS).

As sign languages are non-verbal languages, information is conveyed visually, using a combination of manual and non-manual means of expression. Manual parameters are hand shape, hand posture, hand location, and hand motion. The non-manual parameters include head and body posture, facial expression, gaze and lip patterns.

Non-manual parameters are essential in sign language, since they carry grammatical and prosodic information. Some signs can be distinguished by manual parameters alone, while others remain ambiguous unless additional non-manual information, in particular facial expression, is made available. For instance, the German signs BRUDER (BROTHER) and SCHWESTER (SISTER) are completely identical with respect to gesturing and can only be differentiated by making reference to their lip patterns (Fig. 1).

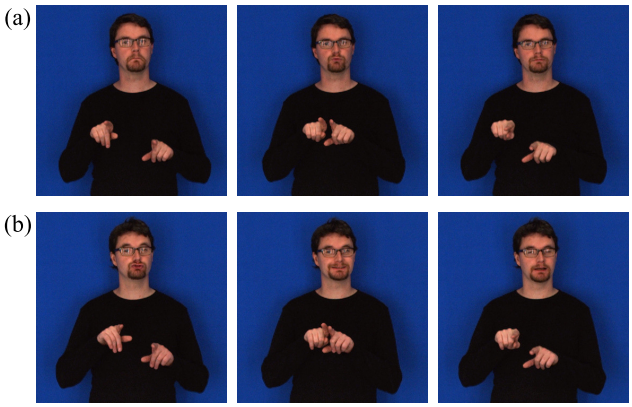

Figure 1. The signs BRUDER (a) and SCHWESTER (b) are identical with respect to manual gesturing but differ in lip patterns.

In the following, some important non-manual parameters will be described in more detail.

Head pose The head pose supports the semantics of sign language. Questions, affirmations, denials, and conditional clauses are communicated, e.g., with the help of the signer's head pose. In addition, information concerning the amount of elapsed time can be encoded as well.

Facial expression Facial expression does not only reflect a person's affect and emotions, but also constitutes a large part of the grammar in sign languages. For example, a change of head pose combined with the lifting of the eye brows corresponds to a subjunctive.

Lip patterns Lip patterns represent the most distinctive non-manual parameter. Certain lip patterns are particular to sign languages whilst others have been borrowed from spoken languages. They solve ambiguities between signs (BROTHER vs. SISTER), specify expressions (MEAT vs. HAMBURGER) and provide information redundant to gesturing to support differentiation of similar signs.

978-1-4244-2154-1/08/$25.00

2. Related Work

The current state in sign language recognition is roughly 30 years behind speech recognition, which corresponds to a gradual transition from isolated to continuous recognition for small vocabulary tasks. Research efforts were mainly focused on robust extraction of manual features or statistical modeling of signs. The reader interested in a detailed survey in sign language recognition is directed to [7].

In general, existing recognition systems can be divided by their means of data acquisition into two groups. Intrusive systems employ data gloves, optical or magnetic markers to determine the signer's manual configuration. For the user, however, this is unnatural and restrictive. In order to overcome this drawback, non-intrusive recognition systems use a video-based approach, which allows extraction of manual and facial features from the same input image. Table 1 lists several publications representing the current state in video-based sign language recognition.

Table 1. Selected video-based sign language recognition systems representing the current state of the art.

Author	Year	Features	Language Level	Language
Vogler [10]	1999	manual	sentence	ASL
Hienz [5]	2000	manual	sentence	DGS
Yang [13]	2002	manual	word	ASL
Zahedi [14]	2007	manual	sentence	ASL
v. Agris [11]	2007	manual	sentence	DGS
Parashar [8]	2003	manual, facial	sentence	ASL

The compilation reveals that most existing recognition systems exploit manual features only; so far facial features were rarely used. Different researchers have recently started to tackle the issue of video-based feature extraction related to non-manual features such as 'head motion' [4] and facial expression [9]. In 2003, Parashar [8] applied a sequential integration approach where the facial information is used to prune the list of word hypotheses generated by manual information. The additional use of information about facial motion increased the accuracy of recognition of continuous words from 88,0% to 92,0%. Furthermore, he was able to detect 'negation' in sentences by means of simple motion trajectory based features 27 out of 30 times. For data acquisition two cameras were used: one for recording the entire signing space and another one focusing on the user's face.

In summary, it can be stated that no publication known, except for [8], has yet proposed combination strategies for manual and non-manual information or proved the expected impact of non-manual features on sign language recognition by presenting evaluation results.

3. System Overview

The following sign language recognition system constitutes the basis for our ongoing research work. A thorough description is given in [6, 12]. Fig. 2 shows a schematic of the underlying concept. The system utilizes a single video camera for data aquisition to ensure user-friendliness. Since sign languages make use of manual and facial means of expression both channels are employed for recognition.

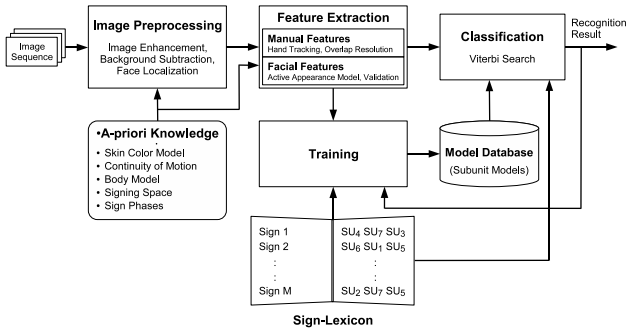

Figure 2. Schematic of the developed sign language recognition system, combining manual and facial features.

For mobile operation in uncontrolled environments sophisticated algorithms were developed that robustly extract manual and facial features. Extraction of manual features relies on a multiple hypotheses tracking approach to resolve ambiguities of hand positions [15]. For facial feature extraction an active appearance model is applied to identify areas of interest such as the eyes and mouth region. Afterwards a numerical description of facial expression and lip outline is computed [2]. Furthermore, the feature extraction stage employs a resolution strategy for dealing with mutual overlapping of the signer's hands and face.

Classification is based on hidden Markov models which are able to compensate time variances in the articulation of a sign. The classification stage is designed for recognition of isolated signs as well as of continuous sign language. For statistical modeling of reference models each sign is represented either as a whole or as a composition of smaller subunits – similar to phonemes in spoken languages [1].

4. Extraction of Manual Features

The feature extraction stage builds on [15] and is designed to process real-world images. It uses a generic skin color model to detect hands and face. The segmentation threshold is automatically chosen so that the resulting face candidate best matches the average face shape. For each pixel, the median color computed from all input images (which are buffered for this purpose) yields a reliable and parameter-free background model. This allows to eliminate static distractors.

The remaining hand candidates still allow many interpretations. Therefore, multiple tracking hypotheses are pursued in parallel. The winner hypothesis is determined only at the end of the sign, using high level knowledge of the human body and the signing process to compute the likelihood of all hypothesized configurations per frame and all transitions between successive frames. This approach exploits all available information for the computation of the final tracking result, yielding robustness and facilitating retrospective error correction.

4.1. Feature Computation

Features are computed from the hand candidate border as shown in Fig. 3. During periods of overlap, template matching is performed to accurately determine the center coordinates x, y using preceding or subsequent unoverlapped views. All other features are linearly interpolated.

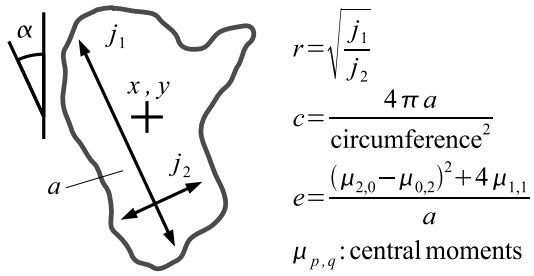

Figure 3. Shape-based features computed for each hand.

Hand center coordinates x, y are specified relative to the corresponding shoulder position, which is estimated from the width w_F and position of the face. For feature normalization all coordinates are divided by w_F, and the area a by its squared value w_F^2. Since $\alpha \in [-90°, 90°)$, it is split into $o_1 = \sin 2\alpha$ and $o_2 = \cos \alpha$ to ensure stability at the interval borders. The features r, c and e describe the shape's axis ratio, compactness, and eccentricity. The derivatives $\dot{x}, \dot{y}, \dot{a}$ complete the 22-dimensional feature vector

$$\boldsymbol{x}_t = [\ \underbrace{x\ \dot{x}\ y\ \dot{y}\ a\ \dot{a}\ o_1\ o_2\ r\ c\ e}_{\text{left hand}}\ \underbrace{x\ \dot{x}\ y\ \dot{y}\ \ldots}_{\text{right hand}}\ ] \quad (1)$$

If the hand is not visible or remains static throughout the sign, its features are set to zero.

5. Extraction of Facial Features

Facial expression analysis and interpretation require that areas of interest, such as the eyes, eyebrows, and mouth (in particular the lips) as well as their spatial relation to each other, have to be extracted from the images first. For this purpose, the face is modeled by an active appearance model (AAM), a statistical model which combines shape and texture information about human faces. Based on an eigenvalue approach the amount of data needed is reduced, hereby enabling real-time processing.

The active appearance model approach is described below in more detail. With regard to the introduced schematic system overview, the localized face region is first cropped and upscaled (Fig. 4, top). Afterwards, AAMs are utilized to match the face graph serving the extraction of facial parameters, such as lip outline, eyes, and brows.

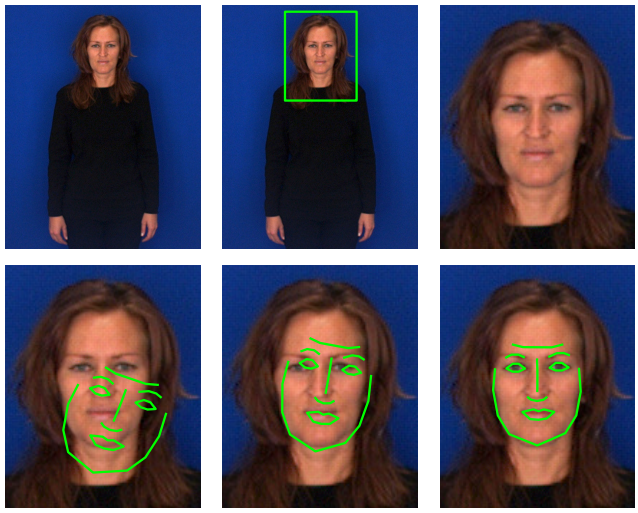

Figure 4. Processing scheme of the face region cropping and the matching of an adaptive face graph.

5.1. Active Appearance Models

Active appearance models contain two main components: a statistical model describing the appearance of an object and an algorithm for matching this model to an example of the object in a new image [3]. In the context of facial analysis, the human face is the object and the AAM can be visualized as a face graph that is iteratively matched to a new face image (Fig. 4, bottom). The statistical models were generated by combining a model of face shape variation with a model of texture variation of a shape-normalised face. Texture denotes the pattern of intensities or colors across an image patch.

Shape model The training set consists of annotated face images where corresponding landmark points have been marked manually on each example. In this framework, the appearance models were trained on face images, each labelled with 50 landmark points at key positions (Fig. 5).

For statistical analysis all shapes must be aligned to the same pose, i.e., the same position, scaling, and rotation. This is performed by a Procrustes analysis which considers the shape in a training set and minimizes the sum of distances with respect to the average shape. After alignment, the shape point sets are adjusted to a common coordinate system.

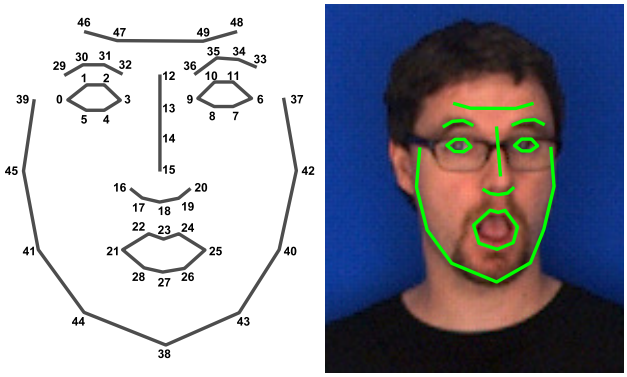

Figure 5. Face graph with 50 landmark points (left) and its application to a specific signer (right).

For dealing with redundancy in high dimensional point sets, AAMs employ a principal component analysis (PCA), a means for dimensionality reduction by first identifying the main axes of a cluster. With the calculated principal components it is possible to reconstruct each sample of the training data. New shape instances can be approximated by deforming the mean shape $\overline{\boldsymbol{x}}$ using a linear combination $\boldsymbol{p}_s$ of the eigenvectors of the covariance matrix $\boldsymbol{\Phi_s}$ as follows

$$\boldsymbol{x} = \overline{\boldsymbol{x}} + \boldsymbol{\Phi_s} \cdot \boldsymbol{p}_s \tag{2}$$

Essentially, the points of the shape are transformed into a modal representation where modes are ordered according to the percentage of variation that they explain. By varying the elements of the shape parameters $\boldsymbol{p}_s$ the shape $\boldsymbol{x}$ may be varied as well.

The eigenvalue λ_i is the variance of the i-th parameter $\boldsymbol{p}_{si}$ over all examples in the training set. Limits are set in order to make sure that a newly generated shape is similar to the training patterns. Empirically, it was found that a maximum deviation for the parameter $\boldsymbol{p}_{si}$ should be no more than $\pm 3\sqrt{\lambda_i}$.

Texture model Data acquisition for shape models is straightforward, since the landmarks in the shape vector constitute the data itself. In the case of texture analysis, one needs a consistent method for collecting the texture information between the landmarks, i.e., an image sampling function needs to be established. Here, a piece-wise affine warp based on the Delaunay triangulation of the mean shape is applied.

Following the warp from an actual shape to the mean shape, a normalization of the texture vector set is performed to avoid the influence from global linear changes in pixel intensities. Hereafter, the analysis is identical to that of the shapes. By applying PCA, a compact representation is derived to deform the texture in a manner similar to what is observed in the training set

$$\boldsymbol{g} = \overline{\boldsymbol{g}} + \boldsymbol{\Phi_t} \cdot \boldsymbol{p}_t \tag{3}$$

where $\overline{\boldsymbol{g}}$ is the mean texture, $\boldsymbol{\Phi_t}$ denotes the eigenvectors of the covariance matrix and finally $\boldsymbol{p}_t$ is the set of texture deformation parameters.

Appearance model The appearance of any example face can thus be summarised by the shape and texture model parameters $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$. In order to remove correlation between both parameters (and to make the model representation even more compact) a further PCA is performed. The combined model obtains the form

$$\boldsymbol{x} = \overline{\boldsymbol{x}} + \boldsymbol{Q_s} \cdot \boldsymbol{c} \tag{4}$$

$$\boldsymbol{g} = \overline{\boldsymbol{g}} + \boldsymbol{Q_t} \cdot \boldsymbol{c} \tag{5}$$

where $\boldsymbol{c}$ is a vector of appearance parameters controlling both shape and texture of the model, and $\boldsymbol{Q_s}$ and $\boldsymbol{Q_t}$ are matrices describing the modes of combined appearance variations in the training set. Fig. 6 presents example appearance models for variations of the first five eigenvectors between $3\sqrt{\lambda}, 0, -3\sqrt{\lambda}$.

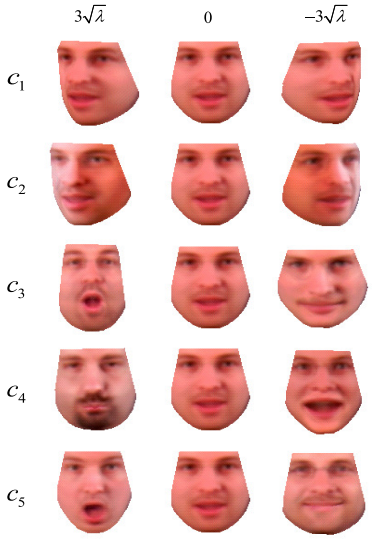

Figure 6. Appearance for variations of the first five eigenvectors c_1, c_2, c_3, c_4, and c_5 between $3\sqrt{\lambda}, 0, -3\sqrt{\lambda}$.

A face can now be synthesized for a given $\boldsymbol{c}$ by generating the shape-free intensity image from the vector $\boldsymbol{g}$ and warping it using the control points described by $\boldsymbol{x}$.

5.2. Feature Computation

After matching the face graph to the signer's face in the input image, areas of interest such as his eyes, eyebrows, and mouth (in particular the lips), as well as their spatial relation to each other, can be easily extracted. Geometric features describing forms and distances serve for encoding the facial expression and the lip outline. They are computed directly from the matched face graph and are divided into three groups (Fig. 7).

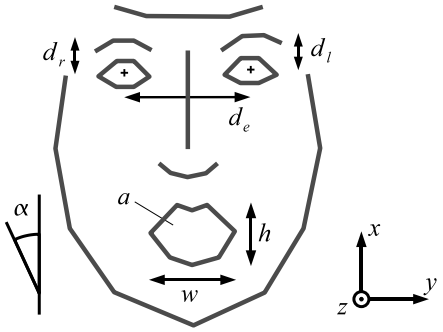

Figure 7. Facial features computed from the matched face graph.

The first group describes the head pose p_x, p_y, p_z in the three dimensional space, while the second group contains the distances d_l, d_r between one eye and its respective eyebrow. In the third group, the lip outline is described by the area a, height h, width w, orientation split into $o_1 = \sin 2\alpha$ and $o_2 = \cos \alpha$, and the shape's axis ratio r as well as the form-features compactness c and eccentricity e. For feature normalization the distances d_l, d_r, h and w are divided by the distance d_e between both eye centers, and the area a by its squared value d_e^2. Finally, the derivatives $\dot{a}, \dot{h}, \dot{w}$ complete the 16-dimensional feature vector

$$\boldsymbol{y}_t = [\ \underbrace{p_x\ p_y\ p_z}_{\text{pose}}\ \underbrace{d_l\ d_r}_{\text{eyebrows}}\ \underbrace{a\ \dot{a}\ h\ \dot{h}\ w\ \dot{w}\ o_1\ o_2\ r\ c\ e}_{\text{lip outline}}\ ] \quad (6)$$

If the face graph cannot be reliably matched, e.g., due to overlapping hands, all facial features are interpolated.

6. Sign Language Corpus

Since we use a vision-based approach for sign language recognition the corpus was recorded on video [11]. In order to facilitate feature extraction recordings are conducted under laboratory conditions, i.e. controlled environment with diffuse lighting and a unicolored blue background. The signers wear dark clothes with long sleeves and perform from a standing position (Fig. 8). A high video resolution of 780×580 pixels at 30 fps ensures reliable extraction of manual and facial features from the same input image.

Figure 8. Example frames taken from the sign language corpus, showing two native signers of different sexes and ages.

The corpus contains videos of isolated signs as well as of continuous sentences. Its vocabulary comprises 450 signs in German Sign Language representing different word types such as nouns, verbs, adjectives, and numbers. Those signs were selected which occur most frequently in everyday conversation and are not dividable into smaller signs. Thus, they are called basic signs in the following.

All basic signs differ in their manual parameters. Many of them, however, change their specific meaning when the manual performance is recombined with a different facial expression. For example, the signs POLITIK (POLITICS) and TECHNIK (ENGINEERING) are identical with respect to gesturing and can only be distinguished by the signer's lip movements. In this case only the former sign is regarded as basic sign, whereas both signs appear in the continuous sentences of the corpus. In total 135 additional signs, derived from the basic signs, were integrated into the database.

Based on this extended vocabulary, overall 780 sentences were constructed (603 for training, 177 for testing). Each sentence ranges from two to eleven signs in length. No intentional pauses are placed between signs within a sentence, but the sentences themselves are separated. All sentences are meaningful and grammatically well-formed. There are no constraints regarding a specific sentence structure.

In order to model interpersonal variance in articulation all 450 basic signs and 780 sentences were performed once by 25 native signers of different sexes and ages. One signer was chosen to be the so-called reference signer. His articulations were recorded not once but even three times.

The corpus will be made available soon for interested researchers in order to establish the first benchmark for signer-independent continuous sign language recognition.

7. Experimental Results

The following experiments were carried out on the sign language corpus described above. Recognition performance for isolated signs was evaluated using the basic signs and for continuous sign language using the sentences. In both cases the evaluation of the signer-dependent (SD) performance is based on the three variations of the reference signer, whereas the signer-independent (SI) recognition rates were determined in a leave-one-out test on all 25 signers. In order to evaluate the performance for different vocabulary sizes, the corpus is divided into three subcorpora simulating a vocabulary of 150, 300, and 450 signs respectively. Table 2 summarizes the experimental results.

All experiments were conducted with three different sets of feature vectors containing: a) only manual features $\boldsymbol{x}_t$, b) only facial features $\boldsymbol{y}_t$, or finally c) a combination of both manual and facial features. In the last case, all features are merged into one feature vector $\boldsymbol{z}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t]$. The obtained results for a) thus represents baselines for c), the additional use of facial features. In all experiments, the classification stage was configured to employ neither subunit models nor any stochastic language model.

Table 2. Signer-independent (SI) recognition of isolated signs and continuous sign language. Recognition rates for signer-dependent (SD) operation are given for comparison.

		Features	Vocabulary Size			
			150	300	450	∅
Isolated	SI	manual	86.7%	83.0%	78.7%	82.8%
		facial	12.2%	10.6%	10.0%	10.9%
		combined	**88.3%**	**84.5%**	**80.2%**	**84.3%**
	SD	manual	95.3%	95.0%	94.4%	94.9%
		facial	48.0%	40.3%	37.1%	41.8%
		combined	**96.0%**	**96.3%**	**96.9%**	**96.4%**
Continuous	SI	manual	62.0%	64.2%	60.6%	62.3%
		facial	8.6%	6.3%	5.4%	6.8%
		combined	**69.0%**	**68.4%**	**65.1%**	**67.5%**
	SD	manual	80.4%	80.6%	80.8%	80.6%
		facial	33.2%	18.5%	12.3%	21.3%
		combined	**87.5%**	**87.4%**	**87.3%**	**87.4%**

When employing manual features only, the recognition rates achieved on the vocabularies presented in Tab. 2 varied between 78.7% and 95.3% for isolated signs and between 60.6% and 80.8% for continous signing. As the production of sign language is subject to high interpersonal variance, signer-independent rates are much lower.

When ignoring manual and using solely facial features, average recognition rates are 26.4% and 14.1% respectively. Hence roughly two of ten signs were recognized just from the signer's face, a result that emphasizes the importance of facial expressions for sign language recognition.

Finally, combining manual and facial features permits to exploit all available information. The recognition rates now ranges from 80.2% to 96.9% in case of isolated signing and from 65.1% to 87.5% in case of continous signing, which is an average improvement of 1.5% and 6.0% respectively compared to using manual features only.

8. Conclusions

In this paper, we described a vision-based sign language recognition system which employs both manual and facial features, extracted from the same input image. For facial feature extraction an active appearance model is applied to identify areas of interest such as the eyes and mouth region. Afterwards a numerical description of facial expression and lip outline is computed. An extensive evaluation proved the importance of integrating facial expressions into automatic sign language recognition. The recognition rates for isolated and continuous signing increased in signer-dependent as well as in signer-independent operation mode. Interestingly, roughly two of ten signs were recognized just from the computed facial features.

Acknowledgments

This work is supported by the Deutsche Forschungsgemeinschaft (German Research Foundation). We thank Uwe Zelle for recording the sign language database.

References

[1] B. Bauer. *Erkennung kontinuierlicher Gebärdensprache mit Untereinheiten-Modellen*. Dissertation, Chair of Technical Computer Science, RWTH Aachen University, 2003.

[2] U. Canzler. *Nicht-intrusive Mimikanalyse*. Dissertation, Chair of Technical Computer Science, RWTH Aachen University, 2005.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. on PAMI*, 23(6):681–685, 2001.

[4] U. M. Erdem and S. Sclaroff. Automatic detection of relevant head gestures in american sign language communication. In *Proceedings of the 16th International Conference on Pattern Recognition*, Quebec, Canada, August 2002.

[5] H. Hienz. *Erkennung kontinuierlicher Gebärdensprache mit Ganzwortmodellen*. Dissertation, Chair of Technical Computer Science, RWTH Aachen University, 2000.

[6] K.-F. Kraiss, editor. *Advanced Man-Machine Interaction*. Springer, 2006.

[7] S. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. on PAMI*, 27(6):873–891, June 2005.

[8] A. Parashar. *Representation and Interpretation of Manual and Non-Manual Information for Automated American Sign Language Recognition*. Master thesis, Department of Computer Science and Engineering, College of Engineering, University of South Florida, 2003.

[9] C. Vogler and S. Goldenstein. Analysis of facial expressions in american sign language. In *Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction*, Las Vegas, USA, July 2005.

[10] C. Vogler and D. Metaxas. Parallel hidden markov models for american sign language recognition. In *Proceedings of the International Conference on Computer Vision*, 1999.

[11] U. von Agris and K.-F. Kraiss. Towards a video corpus for signer-independent continuous sign language recognition. In *Proceedings of the 7th Intl. Workshop on Gesture in Human-Computer Interaction and Simulation*, May 2007.

[12] U. von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss. Recent developments in visual sign language recognition. *Springer Journal on Universal Access in the Information Society*, 6(4):323–362, February 2008.

[13] M.-H. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *IEEE Trans. on PAMI*, 24:1061–1074, 2002.

[14] M. Zahedi. *Robust Appearance-based Sign Language Recognition*. Dissertation, Chair of Computer Science 6, RWTH Aachen, 2007.

[15] J. Zieren and K.-F. Kraiss. Robust person-independent visual sign language recognition. In *Proceedings of the 2nd Iberian Conference on Pattern Recognition and Image Analysis*, Lecture Notes in Computer Science, 2005.