Enhanced Continuous Sign Language Recognition using PCA and Neural Network Features

Yannick L. Gweth, Christian Plahl and Hermann Ney
Chair of Computer Science 6
RWTH Aachen University
Ahorn Str. 55, D-52056 Aachen Germany
{gweth,plahl,ney}@cs.rwth-aachen.de

Abstract

In this work a Gaussian Hidden Markov Model (GHMM) based automatic sign language recognition system is built on the SIGNUM database. The system is trained on appearance-based features as well as on features derived from a multilayer perceptron (MLP). Appearance-based features are directly extracted from the original images without any colored gloves or sensors. The posterior estimates are derived from a neural network. Whereas MLP based features are well-known in speech and optical character recognition, this is the first time that these features are used in a sign language system. The MLP based features improve the word error rate (WER) of the system from 16% to 13% compared to the appearance-based features.

In order to benefit from the different feature types we investigate a combination technique. The models trained on each feature set are combined during the recognition step. By means of the combination technique, we could improve the word error rate of our best system by more than 8% relative and outperform the best published results on this database by about 6% relative.

1. Introduction

Sign language is the main natural communication mean for deaf and hard of hearing people. Sign language is neither international, nor fully based on the local spoken languages. Several different regional languages exist around the word such as American Sign Language (ASL), French Sign Language (LSF) and German Sign Language (DGS). In sign language, the information is conveyed visually and simultaneously using hands, torso and facial expression. Therefore, recognition of sign language is a research area with several challenges in particular in the area of computer vision.

Automatic Sign Language Recognition requires the combined analysis of different information streams including hand gestures (hand shape, orientation and movement trajectories), body poses and facial expressions. Tracking the hands to extract manual features is a challenging task since, the hands move fast, have high degrees of freedom and occlude each other and the face. In some previous works, the signer was therefore required to wear coloured gloves to simplify the tracking and segmentation of the hands [7] [6]. However, that is an unnatural way to sign. A more natural approach uses skin color for segmentation of the hands. Cooper et al. [2] learn a Gaussian skin colour model from face region to detect the hands. In [9], Piater et al. present a hand tracking system based on skin color region segmentation followed by PCA-based template matching. Roussos et al. present in [11] a framework that utilizes novel aspects concerning probabilistic and morphological visual processing for the segmentation, tracking and handshape modeling of the hands. Von Agris et al. use in [14] a generic skin color model and high level knowledge of the human body to detect and segment hands. Starner et al. presented in [12] a video-based approach for recognizing continuous ASL. A single camera is used to extract features as input of an HMM system. Their algorithm scans the image until it finds a pixel of skin color given an a priori model and apply morphological dilatation. They obtain good results with a camera mounted on the desk and in a users cap, but on a small vocabulary of 40 signs only. In [15], Zahedi et al. use different appearance-based features to recognize words of American Sign Language (ASL). Good results are achieved using intensity images, skin color images, and different first- and second-order derivatives. Nevertheless, the database used contains 110 utterances and the vocabulary is limited to 10 words only. All those features have in common that they rely on the segmented input images. Possible segmentation errors will degrade the overall performance of the system. On the other side they have been limited to small lexicons.

978-1-4673-1612-5/12/$31.00 ©2012 IEEE

Facial expressions play a very important role in sign language in resolving ambiguity between signs which have similar manual signs. In [9] and [14] an active appearance model (AAM) [3] is used to combine shape and texture information about face. Fitting errors are a possible error sources since local occlusions can lead the global model to degenerate and lose track of even non-occluded features.

Neural network based features, expecially features trained by multilayer perceptron have become a major component of state-of-the-art speech recognition systems [13, 10]. There, a trained MLP estimates class posterior probabilities which are used as input features to train a GHMM based recognition system. MLP-based posterior probabilities could be used in two different ways within the recognition system. Instead of estimating the Gaussian mixture model the MLP-based posteriors are used directly. The posterior of a GHMM based system are derived by the MLP-based posterior estimates. Therefore, the MLP-based posterior probabilities are divided by the prior state probabilities. This concept is known as the hybrid approach. In the second concept the posterior estimates are used as normal input features to train a GHMM system. This tandem concept has been first published by [5] and is superior to the hybrid approach when a small number of classes is used.

A general review on recent research in sign language and gesture recognition is given by Ong et Ranganath [8]. So far MLP-based features have been employed only for speech recognition and optical character recognition. In speech recognition, they are typically used in combination with standard short-term spectral-based features, and yield consistent improvements in word error rate. In this paper, we investigate the tandem approach and use the MLP-based posteriors features as input for our GHMM system to recognise sign language.

In Section 2 we introduce our GHMM-based automatic sign language recognition system. Followed by the corresponding features used in Section 3. The database used in this work is presented in Section 4. In Section 5, the results of our experiments are presented. We conclude the paper in Section 6.

2. Automatic Sign Language Recognition System Overview

In automatic sign language recognition, we are interested in the best gloss sequence $w_1^N = w_1, ..., w_N$, for which the sequence of observation (i.e. hand patches extracted at position u_1^T from the full image sequence X_1^T) $x_1^T = x_1, ..., x_T$ has been observed (see Figure 1). We want to choose the gloss $\hat{w}_1^N$ that maximizes the posterior probability $p(w_1^N|x_1^T)$ over all possible gloss sequences w_1^N.

$$x_1^T \rightarrow \hat{w}_1^N(x_1^T) = \underset{w_1^N}{\operatorname{argmax}}\{p^\alpha(w_1^N) \cdot p^\beta(x_1^T|w_1^N)\} \quad (1)$$

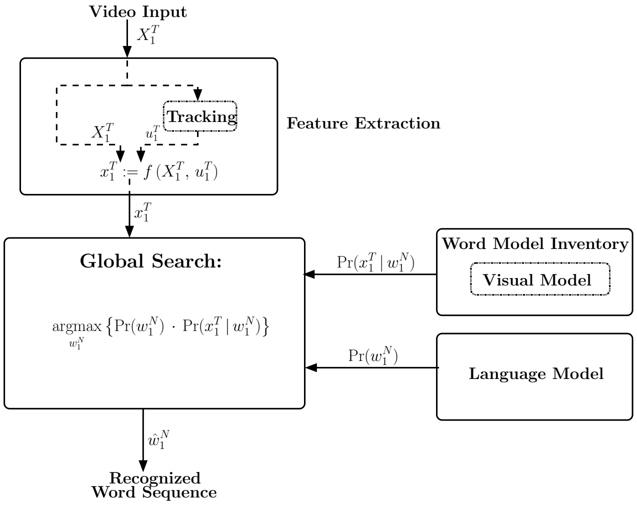


Figure 1. Bayes' decision rule used in ASLR with tracking framework. The result of the tracking is used as input for the recognition system.

where $p(w_1^N)$ is the probability that the gloss sequence w_1^N will be uttered (language model), and $p(x_1^T|w_1^N)$ is the probability of observing features x_1^T given the gloss sequence w_1^N (visual model). α and β are weighting factors for the language model and the visual model. Annotated examples are needed for training the visual model. Each example in the training set is linearly segmented against the models, and the initial estimates of the GHMM parameters are computed. At this point the mapping of each feature vector to the most probably state is conducted. The result of this step is a set of observations pertaining to a specific HMM state, which observations can be used to refine the transition probabilities and also the mean and the covariance matrices associated with this HMM state.

2.1. Synchronous Combination without Retraining

Several separately trained visual models using different feature sets from the same input image can be combined by log-linear combination of the visual probabilities $p_i((x_1^T)_i|w_1^N)$ where w_1^N denotes a sequence of glosses and $(x_1^T)_i$ denotes the sequence of observation extracted using the algorithm i. This approach has been used successfully in automatic speech recognition, leading to significant improvements as presented in [17].

The visual model $p((x_1^T)|w_1^N)$ from Bayes' decision rule in Equation (1) can be redefined as:

$$p((x_1^T)_i|w_1^N) = \prod_i p_i((x_1^T)_i|w_1^N)^{\lambda_i} \quad (2)$$

Consequently, the Bayes' decision rule for log-linear feature combination using a single language model and an acoustic model for each acoustic feature set can be written as:

$$x_1^T \rightarrow \hat{w}_1^N(x_1^T) = \underset{W}{\operatorname{argmax}}\{p(w_1^N)^{\lambda_{lm}} \cdot \prod_i p_i((x_1^T)_i|w_1^N)^{\lambda_i}\} \quad (3)$$

λ_{lm} is the weight of the language model and λ_i the weight of the visual model for the features extracted using the algorithm i. The visual model weights have been optimized empirically.

3. Feature Extraction

3.1. Appearence-based Features

Appearance-based approaches are methods that use the original input images or their projections onto a suitable lower dimensional subspace as features. These features do not rely on any models and have the advantage over invasive system (i.e. where the signer wears gloves or sensors) that simple standard cameras can be used to obtain the images. Different appearance-based features have been used in [15], [16] for recognition of isolated signs in ASL. Although these features have been used for isolated signs, we integrated them into our continuous sign language recognition system. Our baseline system uses therefore appearence-based features from original images (downscaled to 32 × 32 pixels), since they give a global description of manual and non manual features of sign language.

If the resolution of the downscaled images becomes too low, more detailed manual features have to be provided. Our tracking algorithm [4] is used to find the dominant hand (i.e. the hand that is mostly used for one-handed signing). Using the resulting coordinates, we produce 32 × 32 pixels hand patches centered at the tracked position.

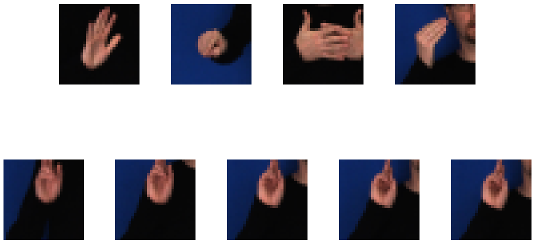
Tracking the hand in an image sequence $X_1^T = X_1, ..., X_T$ is formulated in [4] as a probabilistic optimisation problem. The path of the hand positions $u_1^T = u_1, ..., u_T$ is searched that maximizes the likelihood given the image sequence X_1^T:

$$\hat{u}_1^T = \underset{u_1^T}{\operatorname{argmax}}\{p(u_1^T|X_1^T)\} \quad (4)$$

This approach optimizes over the complete sequences and therefore avoids local decisions that might not be correct. We use cropped hand patches as input features for the system. Examples of cropped hand patches are shown in Table 1. The images represent the hand shapes and orientations of the dominant hand.

Training an ASLR system with high dimensional features requires a huge number of observations to train robust models. Dimensionality reduction technique like *Principal Component Analysis* (PCA) helps to reduce the dimensionality of the resulting feature vectors while rejecting noise and retaining the most relevant information.

Table 1. First row shows hand patches with different hand shapes and the second row patches cropped from a sequence of images. These hand patches are extracted from our tracking framework.



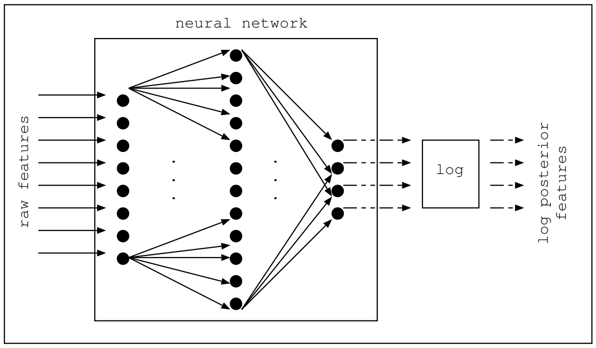


Figure 2. Simple feed-forward neural network. The neural network consist of one hidden layer, one output and one input layer. At the end, the posterior estimates are transformed by logarithm.

3.2. Neural Network based Features

In all our experiments, we have trained a simple feed forward MLP as shown in Figure 2. The MLP consists of one hidden layer, one input layer and one output layer. The two different input feature types used to train the MLP are based on the appearence based features as described in the previous section. The first feature stream consists of the appearance based hand patches of size 32 × 32. These features are concatenated within a sliding window of length 5 and transformed by PCA to a final dimension of 200. The hand-patches are referred to as *RAW* features, whereas the PCA transformed handpatches for MLP training are referred to as *PCA*. The target labels for training the MLP are the 455 different glosses of the system and silence. In the last layer, the softmax activation function is applied to obtain posterior estimates.

In order to include temporal information in the input features for the MLP training, consecutive frames of size 1, 3, 5 or 7 of the appearance based RAW feature set, as well as for the PCA transformed features are concatenated and afterwards normalized by mean and variance. Since the dimension of the hand patches are huge, only one or three consecutive frames have been used as input to train the MLP.

We have trained the MLP with different hidden layer sizes, starting from 1000 hidden nodes up to 2000 hidden nodes. Whereas the best results are achieved with a configuration of 1500 nodes in the hidden layer, the other configuration are only slightly worse. In order to obtain the alignment for training the MLP, the training data has been aligned with a previously trained GHMM based system. This baseline GHMM based system is based on the PCA transformed appearance based hand pathes as described in the previous section.

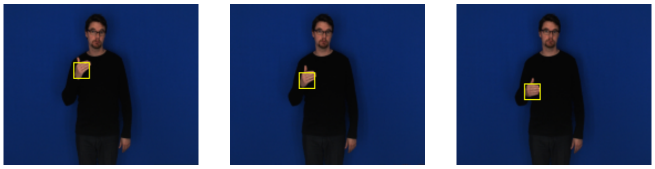
The training of the MLP has been performed on a training set, where the performance of the training is measured on a cross validation set. Therefore, the whole training set has been divided into two disjunct sets. In the training procedure the cross validation set is used to adjust the learning rate and to avoid overfitting of the network. Depending on the feature set and the size of the sliding window the training frame accuracy could be improved from 76% up to 92% on the training set and from 68% up to 75% on the cross validation set. Whereas the best training results has been achieved by the PCA based features and a window size of 7, the best recognition error rate is obtained by another features set.

As described in [5] the final posterior estimates of the MLP are gaussianized. To this purpose, the features are transformed by logarithm. In order to evaluate the performance of the neural network based log posterior features, a GHMM based system has been trained on these log transformed posterior estimates starting from a linear segmentation.

4. SIGNUM Sign Language Corpus

The experiments in this work are performed on the SIGNUM database containing German Sign Language (DGS) [1]. The recordings of the SIGNUM database have been conducted under laboratory condition with uniform background as well as dark clothes for the signer. Table 2 shows some images from this database. The yellow rectangle shows our tracking results of the dominant hand of the signers.

Table 2. Three image examples of the SIGNUM database. Our tracking result of the dominant hand is marked in yellow.



The database is divided into three times 603 sentences for training and three times 177 sentences for testing. It has a basic vocabulary of 450 signs from the DGS which are frequently used in everyday conversation. The basic signs all differ if we consider only the manual information streams. However, some signs change their meaning when they are combined with a different facial expression. The corpus contains 13911 running words divided in 11109 words in the training set and 2802 words in the test set for the signer dependent setup. 3 running words in the test set are not present in the training set (out-of-vocabulary) and can therefore not be recognized. The speaker dependent performance has been evaluated on three recordings of the same signer. The corpus statistics for this task are given in Table 3.

Table 3. Corpus statistics of the SIGNUM database.

	Train	Test
# sentences	1809	531
# frames	416,620	114230
# running glosses	11109	2802
vocabulary size	455	-
# oov rate [%]	-	0.6
perplexity (3-gram LM)	17.9	97.5

5. Experimental Results

5.1. Baseline System

The performance of our system are measured in word error rate (WER). WER is the ratio of insertion, substitution, and deletion of the glosses in the recognized sequence to the total number of signed gloss. Our baseline recognition system is based on appearance based features. Gray level intensities of the image pixels are extracted from the images downscaled to 32 × 32 yielding 1024 dimensional feature vectors. However, high dimensional features are computationally very expensive,but degrade the recognition performance. These problems are resolved by applying PCA for dimensionality reduction. In our experiments 200 dimensions has been empirically proved to be a good size for the resulting feature vectors. Using these features only, we obtain a word error rate of 28.2% on the signer dependent setup of the SIGNUM database. In further experiments, we concatenated several consecutive images into one feature vector before PCA reduction. Table 4 summarizes the results achieved when concatenating 3, 5, 7 images before applying the dimensionality reduction. The best error rate is obtained with 5 images. These results suggest that context information improves the recognition rate.

5.2. Hand Patches Features

Manual components of sign language are mostly conveyed by the dominant hand. Typically, these components are characterized by hand shapes and movements. To introduce this information in the recognition process, we use the dominant hand as described in section 3.1. Table 5 presents the results obtained using those handpatches. We observe

Table 4. Baseline word error rates of our system using PCA-based full image features and concatenation of several images to a feature vector

Features	context	del / ins[%]	WER [%]
Full image	±1	6.7 / 3.0	30.1
	±2	7.2 / 2.3	28.2
	±3	7.1 / 1.8	28.7

an improvement in performance and the error rate drops from 28.2 % to 16.0%. This confirms the assumption that the features of the dominant hand are sufficient features to discriminate most signs in this database. As shown in the previous section, concatenation of consecutive images improves the overall performance.

Table 5. Word error rates of our system using appearence-based handpatch features

Features	context	del / ins[%]	WER [%]
Hand patches	±1	4.6 /1.3	16.6
	±2	2.2 / 3.2	16.0
	±3	- / -	20.8

5.3. Neural Network based Features

In the following experiments we have tested the performance of the MLP-based features. We have used two different features sets, reffered to as RAW and PCA, with several consecutive frames to include temporal dependencies as described in Section 3.2. Experimental results of the PCA features are shown in Table 6 and for the RAW features in Table 7.

The MLP-based posterior estimates based GHMM systems clearly outperform the appearance based systems. Whereas the MLP-based posterior estimates trained on 3 consecutive frames of the RAW feature set already outperform the best previously reported result by more than 3% absolute, the best result is achieved when the PCA transformed features and 3 consecutive images are used as input to train the MLP. Even if the training and cross validation accuracy is increased by using 5 or 7 consecutive frames as input, the corresponding improvement in accuracy do not result in better word error rates in the trained tandem system. Moreover, the results show that the training of the MLPs results in overfitting when the input feature size is increased and the generalization of the MLP-based features gets worse.

Table 6. Error rates of our system with MLP-based features input and different context sizes. The MLP is trained on PCA transformed handpatches.

Features	win	del / ins [%]	WER [%]
PCA	±0	2.2 / 2.3	13.8
	±1	2.0 / 1.1	13.0
	±2	1.7 / 2.6	14.7
	±3	2.1 / 3.8	17.4

Table 7. Error rates of our system using MLP-based features trained on handpatches as input.

Features	win	del / ins [%]	WER [%]
RAW	±0	1.6 / 3.5	14.6
	±1	1.1 / 3.1	13.9

5.4. Synchronous Combination of Features

Three differents features extracted from the original images have been investigated in the previous subsection. The PCA reduced handpatches and MLP-based features derived from the cropped hand patches yield error rates of 16% respectively 13%. The full images downscaled to 32×32 and reduced to 200 components by PCA achieves an error rate of 28%.

A synchronous combination of the three models generated by the different feature groups has been investigated and shows a relative improvement of 7 % compared to the state of the art result of 12.7% published in [14]. Combination results are presented in Table 8. The different feature models have been weighted separately during the recognition stage depending on their significance.

Table 8. Synchronous feature combination without retraining on the SIGNUM dataset

Features	del / ins [%]	WER [%]
Full image (F1)	7.2 / 2.3	28.2
Handpatches (F2)	2.2 / 3.2	16.0
MLP-based posteriors (F3)	2.0 / 1.1	13.0
F1 + F2 + F3	2.1 / 1.5	11.9

6. Conclusion

We presented a sign language recognition system that is able to recognize 84% of sign language sentences on the SIGNUM database using appearance-based features only, and thus avoid unnatural constraints on the signer. This solution is practical because features are extracted directly from the video recorded using a simple camera. This makes the recognition system more practical since signers do not have to wear gloves or markers that make the signing process unnatural. Differents MLP system have been trained using the hand patches and alignments produced by our appearance-base GHMM system. Further posteriors derived from the MLP have been used to train the GHMM system and yield significant improvements. Good results are obtained when combining the advantages of large and nonlinear context modeling via neural networks while profiting from the HMM modeling. The importance of these MLP-based features is also supported by the 19% relative improvement we achieve using the posterior estimates in comparison to the best appearance base feature set.

Different aspects of the signing process are merged together by a synchronous combination without retraining of the individual systems. A suitable combination of the different features at model level yields an improvement of the accuracy and outperforms the best result published on the SIGNUM database by 6% relative.

This work is the first attempt to use neural network based features for continuous sign language recognition. Given this very encouraging start for one signer, we are extending our research to signer independent setup of the SIGNUM database. Further investigations are on the way to evaluate the apport of neural network features on more challenging dataset.

References

[1] V. Agris. Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition. In *Proceedings of GW2007-7th International Workshop on Gesture in Human-Computer Interaction and Simulation 2007–POSTER SESSION*, page 10, 2007.

[2] H. Cooper and R. Bowden. Large lexicon detection of sign language. *Human–Computer Interaction*, pages 88–97, 2007.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, 2001.

[4] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking using dynamic programming for appearance-based sign language recognition. 2006.

[5] H. Hermansky, D. Ellis, and S. Sharma. Tandem connectionist feature stream extraction for conventional HMM systems. pages 1635–1638, 2000.

[6] H. Hienz, B. Bauer, and K. Kraiss. Hmm-based continuous sign language recognition using stochastic grammars. *Gesture-Based Communication in Human-Computer Interaction*, pages 185–196, 1999.

[7] E. Holden and R. Owens. Visual sign language recognition. *Multi-Image Analysis*, pages 270–287, 2001.

[8] S. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891, 2005.

[9] J. Piater, T. Hoyoux, and W. Du. Video analysis for continuous sign language recognition. In *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 22–23, 2010.

[10] C. Plahl, B. Hoffmeister, G. Heigold, J. Lööf, R. Schlüter, and H. Ney. Development of the gale 2008 mandarin lvcsr system. In *Interspeech*, pages 2107–2110, Brighton, UK, Sept. 2009.

[11] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos. Hand tracking and affine shape-appearance handshape subunits in continuous sign language recognition. In *Procs. of Int. Conf. ECCV Wkshp: SGA, Heraklion, Crete*, 2010.

[12] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[13] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. El-Desoky Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney. The rwth 2010 quaero asr evaluation system for english, french, and german. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2212–2215, Prague, Czech Republic, May 2011.

[14] U. von Agris, M. Knorr, and K. Kraiss. The significance of facial features for automatic sign language recognition. In *8th IEEE International Conference on Automatic Face & Gesture Recognition, 2008. FG'08.*, pages 1–6. IEEE, 2008.

[15] M. Zahedi, D. Keysers, and H. Ney. Appearance-based recognition of words in american sign language. *Pattern Recognition and Image Analysis*, pages 511–519, 2005.

[16] M. Zahedi, D. Keysers, and H. Ney. Pronunciation clustering and modeling of variability for appearance-based sign language recognition. *Gesture in Human-Computer Interaction and Simulation*, pages 68–79, 2006.

[17] A. Zolnay, R. Schlüter, and H. Ney. Acoustic feature combination for robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, PA*, 2005.