Speech Recognition Techniques for a Sign Language Recognition System

Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney

Human Language Technology and Pattern Recognition
Computer Science Department 6, RWTH Aachen University, Germany
⟨surname⟩@cs.rwth-aachen.de

Abstract

One of the most significant differences between automatic sign language recognition (ASLR) and automatic speech recognition (ASR) is due to the computer vision problems, whereas the corresponding problems in speech signal processing have been solved due to intensive research in the last 30 years. We present our approach where we start from a large vocabulary speech recognition system to profit from the insights that have been obtained in ASR research.

The system developed is able to recognize sentences of continuous sign language independent of the speaker. The features used are obtained from standard video cameras without any special data acquisition devices. In particular, we focus on feature and model combination techniques applied in ASR, and the usage of pronunciation and language models (LM) in sign language. These techniques can be used for all kind of sign language recognition systems, and for many video analysis problems where the temporal context is important, e.g. for action or gesture recognition.

On a publicly available benchmark database consisting of 201 sentences and 3 signers, we can achieve a 17% WER.

Index Terms: Sign Language Recognition, Video signal processing, Pronunciation Model, Language Model

1. Introduction

Wherever communities of deaf people exist, sign languages develop. As with spoken languages, these vary from region to region and represent complete languages not limited in expressiveness. Linguistic research in sign language has shown that signs mainly consist of four basic manual components [1]: hand configuration, place of articulation, hand movement, and hand orientation. Additionally, non-manual components like facial expression and body posture are used. In continuous sign language recognition, we have to deal with strong coarticulation effects, i.e. the appearance of a sign depends on preceding and succeeding signs, and large inter- and intra-personal variability.

In [2, 3] reviews on recent research in sign language and gesture recognition are presented. In vision-based ASLR, capturing-, tracking- and segmentation problems occur, and it is hard to build a robust recognition framework. Most of the current systems use private databases, specialized hardware [4], and are person dependent [5, 6]. Furthermore, most approaches focus on the recognition of isolated signs only [5, 6], or on the simpler case of gesture recognition [7] for small vocabularies. Our aim is to build a robust, person independent system to recognize sentences of continuous sign language. We use a vision-based approach which does not require special data acquisition devices, e.g. data gloves or motion capturing systems which restrict the natural way of signing. A prototype would just need a simple webcam.

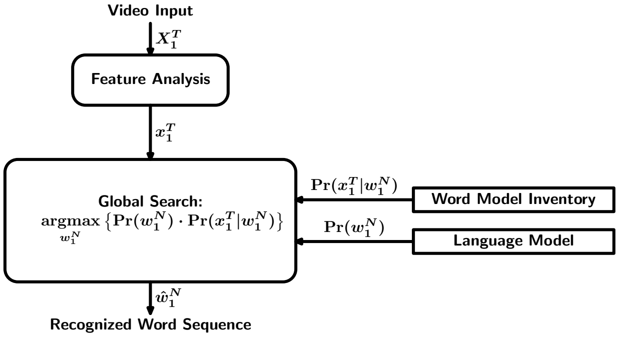

Figure 1: Bayes' decision rule used in ASLR.

Our work is based on a large vocabulary speech recognition system [8]. In particular, we present a complete vision-based framework for person independent continuous sign language recognition as opposed to isolated gesture-recognition works presented by most other authors [5, 6, 7], and analyze the impacts of ASR basic techniques in sign language recognition on a publicly available database with several speakers.

2. System Overview & Features

The ASLR system is based on the Bayes' decision rule. The word sequence which best fits for the current observation to the trained word model inventory (i.e. the acoustic model in ASR) and LM will be the recognition result (see Figure 1).

2.1. Visual Modeling

According to the linguistic work on sign language by Stokoe, a phonological model for sign language can be defined [1], dividing signs in units called "chiremes". However, it is still unclear, how sign language words can be split up into sub-word units (e.g. *phonemes*) suitable for sign language recognition. Therefore, our corpus (c.f. section 3) is annotated in *glosses*, i.e. whole-word transcriptions, and the system is based on whole-word models. Each word model consists of one to three *pseudo-phonemes* modeling the average word length seen in training. Our lexicon defines 247 pseudo-phonemes for 104 words. Each phoneme is modeled by a 3-state left-to-right hidden Markov model (HMM) with three separate Gaussian mixtures (GMM) and a globally pooled covariance matrix.

Due to dialects in natural continuous sign language, signs with the same meaning often differ significantly in their visual appearance and in their duration (e.g. there are 5 different ways to sign the word "bread" in Switzerland). Small differences between the appearance and the length of the utterances are compensated for by the HMMs, but different pronunciations of a sign must be modeled by separate models, i.e. a different number of states and GMMs. Therefore, we added pronunciations to the corpus annotations and adapted our language models (c.f. section 3).

2.2. Language Models

In Bayes' decision rule, the acoustic model (AM) and the language model (LM) have the same impact on the decision, but according to the experience in speech recognition the performance can be greatly improved, if the language model has a greater weight than the acoustic model. The weighting is done by introducing an LM scale α and an AM scale β:

$$\begin{aligned}&\arg\max_{w_1^N}\left\{Pr^{\alpha}(w_1^N)\cdot Pr^{\beta}(x_1^T|w_1^N)\right\}\\ &=\arg\max_{w_1^N}\left\{\frac{\alpha}{\beta}\log Pr(w_1^N)+\log Pr(x_1^T|w_1^N)\right\}\end{aligned}$$

The factor $\frac{\alpha}{\beta}$ is referred to as *language model factor*. A trigram LM was trained using the SRILM toolkit with modified Kneser-Ney discounting with interpolation.

2.3. Appearance-Based Features

In our baseline system we use appearance-based image features only, i.e. thumbnails of video sequence frames. These intensity images scaled to 32×32 pixels serve as good basic features for many image recognition problems, and have already been successfully used for gesture recognition [9]. They give a global description of all (manual and non-manual) features proposed in linguistic research. language. The baseline system is Viterbi trained and uses a trigram LM (c.f. subsection 2.2). In subsequent steps, this baseline system is extended by features accounting for the hands and their positions.

2.4. Manual Features

To extract manual features, the dominant hand (i.e. the hand that is mostly used for one-handed signs such as finger spelling) is tracked in each image sequence. Therefore, a robust tracking algorithm for hand and head tracking is required as the signing hand frequently moves in front of the face, may temporarily disappear, or cross the other hand. Instead of requiring a near perfect segmentation for these body parts, the decision process for candidate regions is postponed to the end of the entire sequences by tracing back the best decisions [10]. Given the hand position (HP) $u_t = (x, y)$ at time t in signing space, features such as hand velocity (HV) $m_t = u_t - u_{t-\delta}$ can easily be extracted.

The hand trajectory (HT) features presented here are similar to the features presented in [5]. Here we calculate global features describing geometric properties of the hand trajectory in a certain time window $2\delta + 1$ around time t by an estimation of the covariance matrix

$$\Sigma_t = \frac{1}{2\delta+1}\sum_{t'=t-\delta}^{t+\delta}(u_{t'}-\mu_t)(u_{t'}-\mu_t)^T$$

and $\mu_t = \frac{1}{2\delta+1}\sum_{t'=t-\delta}^{t+\delta} u_{t'}$. For $\Sigma_t \cdot v_{t,i} = \lambda_{t,i} \cdot v_{t,i}, i \in \{1, 2\}$, the eigenvalues $\lambda_{t,i}$ and eigenvectors $v_{t,i}$ of the covariance matrix can then be used as global features, describing the form of the movement. If one eigenvalue is significantly larger than the other, the movements fits a line, otherwise it is rather elliptical. The eigenvector with the larger corresponding eigenvalue can be interpreted as the main direction of the movement.

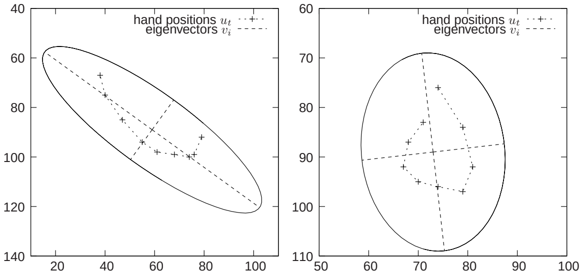

Figure 2: Examples of different hand trajectories and corresponding eigenvectors for $\delta = 4$. The covariance matrices are visualized as ellipses with axes of length $\sqrt{\lambda_i}$.

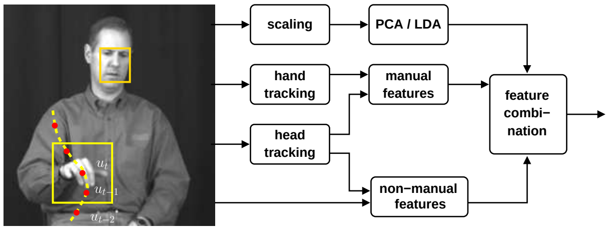

Figure 3: Composite Features using speech signal processing network.

Figure 2 shows some examples of trajectories and their eigenvectors and eigenvalues.

2.5. Feature Selection & Combination

A known problem with appearance-based features are border pixels that do not help in the classification and have very low variance. To resolve this problem, dimensionality reduction techniques like PCA or LDA are commonly applied. LDA is often used in speech recognition to combine and reduce features while maximizing the linear separability of the classes in the transformed feature space. Furthermore in ASR, succeeding feature vectors are commonly concatenated before the LDA transformation is applied to account for temporal dependencies. A critical parameter is the number of succeeding feature vectors that are concatenated, because for a growing window size an increasing amount of training data is needed.

Figure 3 shows how we extract and combine features. The results achieved using different features and combination methods are presented in section 3.

3. Experimental Results

To tune and test our system, we assembled the RWTH-Boston-104 corpus[1] as a subset of a much larger database of sign language sentences that were recorded at Boston University for linguistic research [11]. The RWTH-Boston-104 corpus consists of 201 sequences, and the vocabulary contains 104 words. The sentences were signed by 3 speakers (2 female, 1 male) and the corpus is split into 161 training and 40 test sequences. An overview on the corpus is given in Table 1: 26% of the training data are singletons, i.e. a "one-shot training" occurs. The sentences have a rather simple structure and therefore the language model perplexity (PP) is low. The test corpus has one out-of-vocabulary (OOV) word. Obviously, this word cannot be recognized correctly.

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

Table 1: RWTH-Boston-104 corpus statistics

	Training	Test
sentences	161	40
running words	710	178
vocabulary	103	65
singletons	27	9
OOV	-	1

LM type	PP
zerogram	106.0
unigram	36.8
bigram	6.7
trigram	4.7

Table 2: Baseline results using appearance-based features

Features	Dim.	[%WER]
intensity (w/o pronunciations)	1024	54.0
intensity (w/ pronunciations)	1024	37.0
intensity (w/ pronunciations + tangent distance)	1024	33.7
motion (pixel based)	1024	51.1
intensity+motion	2048	42.1

The HMM based ASR framework offers various tuning possibilities. From former experiments we know that a high number of states per word and a high number of mixture densities have a positive impact on the recognition performance.

Baseline. First, we analyze different appearance-based features for our baseline system. Table 2 gives an overview of results obtained with the baseline system for a few different features. It can be seen that intensity images compared with tangent distance [9] already lead to reasonable results. Contrary to ASR, the first-order time derivatives of the intensity features (i.e. the motion feature) or the concatenation of them with the intensity features (i.e. the intensity+motion feature) usually do not improve the results in video analysis, as the time resolution is much lower (e.g. 25 or 30 video frames/sec compared to 100 acoustic samples/sec in speech). The simplest and best appearance-based feature is to use intensity images down scaled to 32×32 pixels. This size, which was tuned on the test set, was reported to also work reasonably well in previous works [9, 12]. Another important point is the usage of pronunciation modelling in sign language: it can be seen that by adding pronunciations to the corpus and the adaptation of the used trigram language model, the system can already be improved from 54.0% to 37.0% WER.

Feature Reduction. Obviously, the high dimensional appearance-based feature vectors encode a lot of background noise and one would need many more observations to train a robust model. To reduce the number of features and noise and thus the number of parameters to be learned in the models, we apply linear feature reduction techniques to the data. The best obtained result with LDA is 36% WER, whereas with PCA a WER of 27.5% can be obtained. Although theoreticaly LDA should be better suited for pattern recognition tasks, here the training data is insufficient for a numerically stable estimation of the LDA transformation and thus PCA, which is reported to be more stable for high dimensional data with small training sets outperforms LDA [12].

Windowing. We experimentally evaluated the incorporation of temporal context by concatenating features $x_{t-\delta}^{t+\delta}$ within a sliding window of size $2\delta+1$ into a larger feature vector $\hat{x}_t$ and then applying linear dimensionality reduction techniques as in ASR to find a good linear combination of succeeding feature vectors. The outcomes of these experiments are given in Figure 4 and Figure 5 and again, the PCA outperforms the LDA. The best result (21.9% WER) is achieved by concatenating and reducing five PCA transformed (i.e. a total of 110×5 components) frames to 100 coefficients, whereas the best result obtained with LDA is only 25.8% WER, probably again due to insufficient training data (as 2 matrices have to be estimated for LDA). Furthermore, windowing with large temporal contexts increases the system performance, as coarticulation effects are described now.

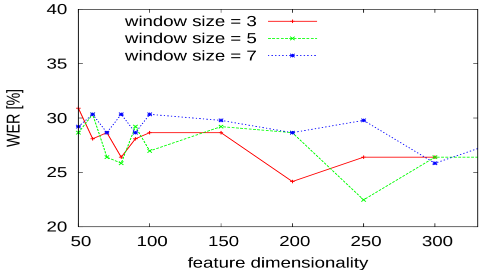

Figure 4: Combination of PCA-frames using LDA windowing

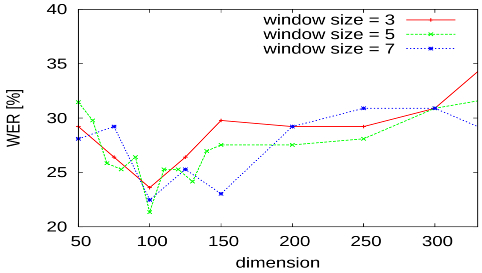

Figure 5: Combination of PCA-frames using PCA windowing

Feature and Model Combination. As explained before, in sign language, different channels have to be considered. To incorporate the data from these different channels, we propose to use a combination of features. Results for various combinations are presented in Table 3 and a clear improvement can be observed. Many other feature combinations are possible and were tested, but as we do not want to overfit our system, we just extracted the manual features from the dominant-hand related to linguistic research (i.e. place of articulation, hand movement, and hand orientation. The hand configuration is encoded in the complete PCA-frame).

A log-linear combination of two independently trained models (windowed PCA-frame+HT and windowed PCA-frame+HV) leads to a further improvement. A WER of 17.9% is achieved (i.e. 17 del., 3 ins., and 12 subst.), where the model weights have been optimized empirically. This is in accordance to experiments in other domains where the combination of different models leads to an improvement over the individual models [13]. In this case, the improvement is due to a better performance of the HT feature for long words and a better performance of the HV feature for short words. A combination on the feature level cannot exploit this advantage because only one alignment is created where the combination of two separately trained models profits from *two independent alignments*, one performing well for long words and the other performing well for short words. Note that the HT feature is strongly distorted for short words (i.e. less than 5 states) because at the word boundaries strong coarticulation effects occur.

Language Model. Figure 6 shows the effect of using different n-gram language models and scales. As in ASR, the language model adaptation by using sign language pronunciations achieves large improvements (c.f. baseline results). Interestingly, the improvement factors achieved are similar to those from speech recognition [14]. Due to the lack of training data for the LM no further improvements are expected for e.g. 4-gram language models. It can also be seen that the LM scale is one of the most important parameters of a continuous sign language recognition system.

Table 3: Results for feature combinations with hand features

Features	Dimensionality	[% WER]
PCA-frame	110	27.5
PCA-frame, hand-position (HP)	112	25.3
PCA-frame, hand-velocity (HV)	112	24.2
PCA-frame, hand-trajectory (HT)	112	23.6
model-combination	2×100	17.9

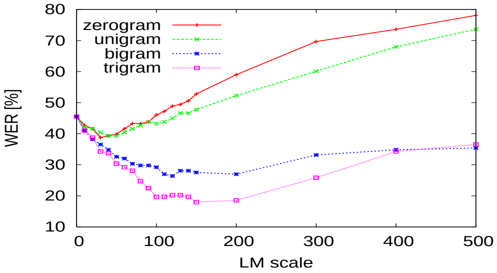

Figure 6: Results for different LMs and scales

4. Summary & Conclusion

We presented a vision-based approach to continuous automatic sign language recognition. We have shown that appearance-based features, which have been proven to be a powerful tool in many image recognition problems, are also well suited for the recognition of sign language. Furthermore, we have shown that many of the principles known from ASR, such as pronunciation and language modelling can directly be transfered to the new domain of vision-based continuous ASLR. We presented very promising results on a publicly available benchmark database of several speakers which has been recorded without any special data acquisition tools.

Combining different data sources, suitable language and pronunciation modelling, temporal contexts, and model combination, the 37% WER of our baseline system could be improved to 17.9% WER. The results suggest that for high dimensional data and the relatively low amount of available training data, PCA outperforms LDA for this task and that context information is as important as it is in ASR.

Outlook. Obviously, a large amount of work still needs to be done for the vision part of the system. New features describing the hand and body configuration as e.g. in [15] should be analyzed and combined with the existing feature set. Certainly an important step is the definition of sub-word units which would allow recognition with a larger vocabulary and the consideration of context dependency with suitable models for coarticulation. Great improvements are also expected from speaker adaptation techniques such as MLLR, because of the large interpersonal differences in sign language.

In order to build a vision-based speech-to-speech system for deaf people, our system is connected to a statistical machine translation system. In preliminary translation experiments presented in [16], the incorporation of the tracking data for the deixis words helped the translation system to discriminate between deixis as distinctive article, locative or discourse entity reference function. Furthermore we collected a new publicly available sign language database with currently 843 sentences, a vocabulary of 403 words (482 words w/ pronunciations), which was signed by 5 speakers.

5. References

[1] W. Stokoe, D. Casterline, and C. Croneberg, *A Dictionary of American Sign Language on Linguistic Principles*, Gallaudet College Press, Washington D.C., USA, 1965.

[2] S. Ong and S. Ranganath, “Automatic sign language analysis: A survey and the future beyond lexical meaning,” *IEEE Trans. PAMI*, vol. 27, no. 6, pp. 873–891, June 2005.

[3] T.S. Huang Y. Wu, “Vision-based gesture recognition: a review,” in *Gesture Workshop*, Gif-sur-Yvette, France, Mar. 1999, vol. 1739 of *LNCS*, pp. 103–115.

[4] G. Yao, H. Yao, X. Liu, and F. Jiang, “Real time large vocabulary continuous sign language recognition based on op/viterbi algorithm,” in *Intl. Conf. Pattern Recognition*, Hong Kong, Aug. 2006, vol. 3, pp. 312–315.

[5] C. Vogler and D. Metaxas, “A framework for recognizing the simultaneous aspects of american sign language,” *Computer Vision & Image Understanding*, vol. 81, no. 3, pp. 358–384, Mar. 2001.

[6] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, “A linguistic feature vector for the visual interpretation of sign language,” in *European Conf. Computer Vision*, 2004, vol. 1, pp. 390–401.

[7] S. B. Wang, A. Quattoni, Louis-Philippe Morency, David Demirdjian, and Trevor Darrell, “Hidden conditional random fields for gesture recognition,” in *Computer Vision & Pattern Recognition*, New York, USA, June 2006, vol. 2, pp. 1521–1527.

[8] J. Lööf, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schluter, and H. Ney, “The 2006 RWTH parliamentary speeches transcription system,” in *ICSLP*, Pittsburgh, PA, USA, Sept. 2006.

[9] D. Keysers, T. Deselaers, C. Gollan, and H. Ney, “Deformation models for image recognition,” *IEEE Trans. PAMI*, p. to appear, 2007.

[10] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney, “Tracking using dynamic programming for appearance-based sign language recognition,” in *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, Southampton, Apr. 2006, pp. 293–298.

[11] C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee, *The Syntax of American Sign Language*, MIT Press, 1999.

[12] T. Kölsch, D. Keysers, H. Ney, and R. Paredes, “Enhancements for local feature based image classification,” in *Intl. Conf. Pattern Recognition*, Cambridge, UK, Aug. 2004, vol. 1, pp. 248–251.

[13] A. Zolnay, R. Schlüter, and H. Ney, “Acoustic feature combination for robust speech recognition,” in *ICASSP*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 457–460.

[14] D. Klakow and J. Peters, “Testing the correlation of word error rate and perplexity,” *Speech Communication*, vol. 38, pp. 19–28, 2002.

[15] A. Agarwal and B. Triggs, “Recovering 3d human pose from monocular images,” *IEEE Trans. PAMI*, vol. 28, no. 1, pp. 44–58, Jan. 2006.

[16] P. Dreuw, D. Stein, and H. Ney, “Enhancing a sign language translation system with vision-based features,” in *Intl. Workshop on Gesture in HCI and Simulation 2007*, Lisbon, Portugal, May 2007, p. to appear.