Visual Modeling and Feature Adaptation in Sign Language Recognition

Philippe Dreuw and Hermann Ney

Human Language Technology and Pattern Recognition, RWTH Aachen University, D-52056 Aachen
E-Mail: {dreuw,ney}@cs.rwth-aachen.de
Web: http://www-i6.informatik.rwth-aachen.de

Abstract

We propose a tracking adaptation to recover from early tracking errors in sign language recognition by optimizing the obtained tracking paths w.r.t. the hypothesized word sequences of an automatic sign language recognition system. Hand or head tracking is usually only optimized according to a tracking criterion. As a consequence, methods which depend on accurate detection and tracking of body parts lead to recognition errors in gesture and sign language processing. Similar to speaker dependent feature adaptation methods in automatic speech recognition, we propose an automatic visual alignment of signers for vision-based sign language recognition. Furthermore, the generation of additional virtual training samples is proposed to reduce the lack of data problem in sign language processing, which often leads to "one-shot" trained models. Most state-of-the-art systems are speaker dependent, and consider tracking as a preprocessing feature extraction part. Experiments on a publicly available benchmark database show that the proposed methods strongly improve the recognition accuracy of the system.

1 Introduction

Phonological analysis going back to Stokoe et al [6] has revealed that signs are made up out of basic articulatory units, initially referred to as cheremes by Stokoe, now commonly called *phonemes* because of their similarity with the discriminatory units that compose words in spoken languages.

Signs are generally decomposed analytically for purposes of linguistic analysis into hand shape, orientation, place of articulation, and movement (with important linguistic information also conveyed through non-manual gestures, i.e., facial expressions and head movements). Main differences between spoken language and sign language are due to language characteristics like simultaneous facial and hand expressions, references in the virtual signing space, and grammatical differences as explained in the following paragraphs.

Simultaneousness: One major issue in sign language recognition compared to speech recognition is the possible simultaneousness [8]: a signer can use different communication channels (facial expression, hand movement, and body posture) in parallel.

Signing Space: Entities like persons or objects can be stored in the sign language space, i.e. the 3D body-centered space around the signer, by executing them at a certain location and later just referencing them by pointing to the space [9]. A challenging task is to define a model for spatial information containing the entities created during the sign language discourse.

Environment: Further difficulties for such sign language recognition frameworks arise due to different environment assumptions. Most of the methods developed assume closed-world scenarios, e.g. simple backgrounds, special hardware like data gloves, limited sets of actions, and a limited number of signers, resulting in different problems in sign language feature extraction.

Speakers and Dialects: As in automatic speech recognition we want to build a robust, person-independent system being able to cope with different dialects. Speaker adaptation techniques known from speech recognition can be used to make the system more robust. While for the recognition of signs of a single speaker only the intrapersonal variabilities in appearance and velocity have to be modelled, the amount and diversity of the variabilities is enormously increased with an increasing number of speakers.

Coarticulation and Epenthesis: In continuous sign language recognition, as well as in speech recognition, coarticulation effects have to be considered. Furthermore, due to location changes in the virtual signing space, we have to deal with the movement epenthesis problem [8, 10]. Movement epenthesis refers to movements which occur regularly in natural sign language in order to change the location in signing space.

Silence: As opposed to automatic speech recognition, where usually the energy of the audio signal is used for the silence detection in the sentences, new features and models will have to be defined for silence detection in sign language recognition. Silence cannot be detected by simply analyzing motion in the video, because words can be signed by just holding a particular posture in the signing space. A thorough analysis and a reliable detection of silence in general and sentence boundaries in particular are important to reliable speed up and automate the training process in order to improve the recognition performance.

2 Visual Modeling

However, it is still unclear how best to approach recognition of these articulatory parameters. Although phonemes in spoken language are sequential, notwithstanding coarticulation effects, in signed languages phonemes are realized simultaneously. The hand is simultaneously in a particular configuration, orientation, and location as it undergoes movement. The recognition of (linguistic) phonemes could be possible in a multi-channel approach, where the correct and combined alignment of the independent systems remains a challenge. Here, we focus on the recognition of *glosses* in the annotations, i.e., whole-word transcriptions, and the system is based on whole-word models. Each word model consists of a temporal division into one to three *pseudo-phonemes* modeling the average word length seen in training. Each pseudo-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.

Figure 1: Examples of different hand patches extracted from tracking framework with their corresponding back-projections from PCA space using a 1600×30 dimensional PCA matrix

2.1 Recognition System: Overview

In a vision-based system, tracking-based features have to be extracted at unknown positions $u_1,\ldots,u_T$ in a video sequence of images $X_1,\ldots,X_T$, with $u=(x,y)$ a 2D tracking position in the image.

In an automatic sign language recognition (ASLR) system for continuous sign language, we are searching for an unknown word sequence w_1^N, for which the sequence of features $x_1^T = f(X_1^T, u_1^T)$ best fits to the trained models.

Opposed to the recognition of dynamic (but isolated) gestures, we maximize the posteriori probability $\Pr(w_1^N|x_1^T)$ over all possible word sequences w_1^N with unknown number of words N. This is modeled by Bayes' decision rule. The optimal word sequence is found using maximum approximation over all possible hidden Markov model (HMM) temporal state sequences:

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}} \left\{ p(w_1^N) \max_{s_1^T} \prod_{t=1}^{T} \left\{ p(f(X_t,u_t)|s_t,w_1^N) \cdot p(s_t|s_{t-1},w_1^N) \right\} \right\} \quad (1)$$

where $\Pr(w_1^N)$ is the a-priori probability for the word sequence w_1^N given by the language model (LM). Here, the system is Viterbi trained and uses a smoothed trigram language model. $\Pr(x_1^T|w_1^N)$ is the probability of observing features x_1^T given the word sequence w_1^N, referred to as visual model (VM), with $x_1^T := f(X_1^T, u_1^T)$ hand tracking features extracted at positions u_1^T in the image observation sequence X_1^T.

2.2 Features

We use appearance-based image and hand features, i.e. thumbnails of video sequence frames, which can be reduced by linear feature reduction methods like PCA or LDA. These features give a global description of all (manual and non-manual) features that have been shown to be linguistically important.

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 [2]. Given the hand position (HP) $u_t=(x,y)$ at time t in signing space, features such as hand velocity (HV) $m_t = u_t - u_{t-1}$ or hand trajectory (HT) can easily be extracted [3]. Furthermore, the obtained tracking rectangle Q_t, which is the set of pixel coordinates around a hand tracking center u_t of size $w\times h$, can be used to extract appearance-based hand features (see Figure 1).

We focus in the following sections only on these rather low-level frame and hand based features instead of possibly high-level features presented e.g. in [5]. Nevertheless, the achieved results in section 4 even outperform other approaches on the same benchmark set.

2.3 Visual Speaker Alignment

Due to the usage of appearance-based features (c.f. subsection 2.2), which do not require a near perfect segmentation of the body parts and encode also the scales and relative depth of the objects to be extracted (e.g. hands and head), the bodies of the signing speakers should have the same baseline depth. With an increasing number of signers in the database, the variability in data, such as different head or hand sizes due to different speaker sizes, has to be modeled.

Similar to a speaker dependent feature adaptation in ASR, we propose to adapt the vision-based features, too. Based on a face detection in a reference recording of each speaker with the Viola & Jones [7] head detection method, we propose to automatically align the speakers.

For every speaker n in a set of $1,..,N$ speakers, we want to find the speaker dependent affine warping matrix A_n, so that the difference between all overlapping speaker images, i.e. the cropped regions-of-interest (ROI), and their corresponding detected heads is minimal.

Similar to a tracking rectangle Q_t let Q_n be now the set of pixel coordinates around a ROI center u_n of size $w\times h$. This means that for every pixel position $u=(x,y)$ in a ROI Q_n, we want to optimize the parameters of an affine 2×3 warping matrix

$$A_n = \begin{bmatrix} A_{11} & A_{12} & b_1 \\ A_{21} & A_{22} & b_2 \end{bmatrix}$$

with

$$Q'_n = \{u_t + u : u \in QA_n\}$$
$$QA_n = \{(A_{11}i + A_{12}j + b_1, A_{21}i + A_{22}j + b_2), (i,j) \in Q\}$$

such that the difference between the warped ROI Q'_m and the warped target ROI Q'_n is minimal.

Based on the ROI Q_n and the face detection rectangle $r_n(x,y,w,h) := \{(x-w/2, y-h/2), (x+w/2, y+h/2)\}$ of a target speaker n ($n \in 1,..,N$), the speaker dependent affine warping matrices A_m of the remaining $N-1$ speakers are optimized w.r.t. the difference between the ROIs Q'_n and Q'_m and a face penalty function which penalizes large differences between face position and ratio:

$$q(r_nA_n, r_mA_m) = \sqrt{(x'_{r_n} - x'_{r_m})^2 + (y'_{r_n} - y'_{r_m})^2} + (w'_{r_n} - w'_{r_m})^2 + (h'_{r_n} - h'_{r_m})^2 \quad (2)$$

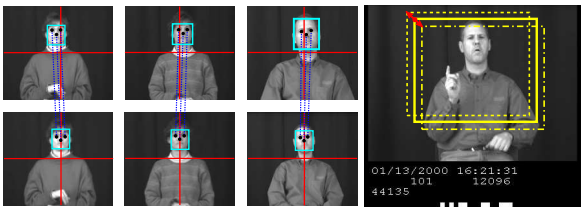
with $r'_n := r_nA_n = \{(A_{11}i + A_{12}j + b_1, A_{21}i + A_{22}j + b_2), (i,j) \in r_n\}, n = 1,..,N$, the affine transformed face rectangle.

For an appearance-invariant (e.g. background or clothing) matching score of the speakers, the gray intensity images X_n are thresholded to binary images (denoted by $\tilde{X}_n$), but any other pre-processing could be used here.

This visual speaker alignment (VSA) can then be expressed with the following optimization criterion:

$$\min_{A_n,A_m} \left\{ \sum_{\substack{u\in Q'_n \\ u'\in Q'_m}} (\tilde{X}_n[u] - \tilde{X}_m[u'])^2 + \alpha \cdot q(r'_n, r'_m) \right\} \quad (3)$$

To speed up the VSA optimization, all ROIs are first horizontally translated to center the speakers' head to optimize the warping matrices only w.r.t. vertical translation



(a) (b)

Figure 2: Automatic speaker alignment based on face detection (a) and virtual training samples generation (b) by slightly distorted cropping positions.

and scaling (rotation could also be considered if necessary). Figure 2 gives an overview of the automatic speaker alignment and virtual training data generation. Figure 2 (a) shows the resulting speaker aligned ROIs cropped from the orginal frames in Figure 2 (b).

2.4 Virtual Training

In order to build a robust recognition system which can recognize continuous sign language speaker independently, we have to cope with various difficulties: (i) *coarticulation:* the appearance of a sign depends on the preceding and succeeding signs. (ii) *inter- and intrapersonal variability:* the appearance of a particular sign can vary significantly in different utterances of the same signer and in utterances of different signers. To model all these variabilities, a large amount of training data is necessary to estimate the parameters of the system reliably.

Due to the lack of data in video benchmark databases for sign language recognition, some visual models contain only a few observations per density. Even "one-shot" training is necessary for singletons (c.f. section 4). This results in too sharp means which do not generalize well on unseen data.

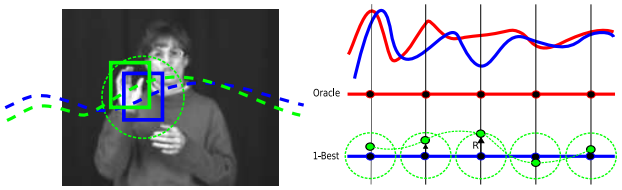
However, for other pattern recognition problems it has been reported that the usage of additional virtual training samples can significantly improve the system performance [1]. Here, as only a region-of-interest (ROI) is cropped from the original video frames, the amount of training data can be increased by virtual training samples, i.e. ROIs extracted at slightly shifted positions from the original ROI position. The cropping position, i.e. the ROI center (x,y), is shifted by $\pm\delta$ pixels in x- and y-direction. For $\delta = 1$, the training corpus is already enlarged by a factor of nine.

The proposed virtual training samples generation can be interpreted as distortion and adaptation on the signal level. Each additional virtual training sample may lead to a slightly different tracking path and thus effectively different tracking paths are considered in training and testing.

3 Model-Based Tracking Path Adaptation

Most state-of-the-art systems consider tracking as a preprocessing feature extraction part, where tracking errors lead to recognition errors. Therefore, we propose to adapt the tracking path to the hypothesized word sequence by a locally distortion within a range R of the given tracking path.

Usually the local adaptation search is chosen very small which results in smooth paths and better hand hypotheses matching to the visual models resulting in better emission scores (see Figure 3).



(a) (b)

Figure 3: Path adaptation: distorted hand hypotheses can be weighted by the distance to the given tracking path (a). The path optimal w.r.t. the tracking criterion (blue line) can be distorted locally (b) during the search in order to obtain tracking adapted features being optimal w.r.t. the hypothesized word sequence.

Furthermore, it is possible to penalize locations far away from the original tracking path. Each distortion depends on the currently hypothesized word (i.e. the trained hand models), which changes the visual model probability in Equation 1 as follows: $\Pr(x_1^T, s_1^T | w_1^N) =$

$$\prod_{t=1}^{T} \Bigg\{ \max_{\substack{\delta\in\{(x,y):\\ -R\le x,y\le R\}}} \{p(\delta)\cdot p(f(X_t,u_t+\delta)|s_t,w_1^N)\} \cdot p(s_t|s_{t-1},w_1^N) \Bigg\} \quad \text{with} \quad p(\delta) = \frac{\exp(-\delta^2)}{\exp(\sum_{\delta'} -\delta'^2)}.$$

The path distortion model prunes the search space starting from a path being optimal to a tracking criterion in order to obtain a distorted path according to the hypothesized word sequence. Here, we assume that the tracking may be inaccurate up to $\pm\delta$ pixels and allow for compensating tracking errors up to this range in the recognition phase.

4 Experimental Results

For our experiments, we use a publicly available database of 201 American Sign Language sentences performed by 3 different signers, 161 are used for training and 40 for testing [3]. On the average, these sentences consist of 5 words out of a vocabulary of 104 unique words. The test corpus has one out-of-vocabulary (OOV) word. These words are not included in the recognition vocabulary and thus cannot be recognized. Every OOV word leads to at least one recognition error. 26% of the vocabulary words seen in training are singletons (i.e. words which occur only once in the training corpus).

We use only unseen data from the test sentences for evaluation. As we are dealing with continuous sign language sentences (instead of isolated gestures only), the recognition experiments are evaluated using the word error rate (WER) in the same way as it is done in speech recognition.

In order to analyze the proposed tracking rescoring and adaptation methods, here we focus only on the usage of appearance-based hand features in contrast to full appearance-based frame features, more complex tracking features, and their combinations as proposed by the authors of [3].

The results using the model-based tracking path adaptation in combination with a visual speaker alignment

Table 1: Rescoring results using the path distortion model for visual speaker alignment and virtual training samples.

Features / Adaptation	WER[%]			
	Baseline	VSA	VTS	VSA+VTS
Frame 32×32	35.62	33.15	27.53	**24.72**
PCA-Frame (200)	30.34	27.53	19.10	**17.98**
Hand (32×32)	45.51	33.15	20.79	21.91
+ distortion ($R = 10$)	41.03	29.78	16.29	16.85
+ δ-penalty	35.96	26.40	**15.73**	16.85
PCA-Hand (30)	28.65	33.15	23.60	24.16
+ distortion ($R = 10$)	33.15	28.65	21.35	16.85
+ δ-penalty	29.78	24.72	17.98	**16.29**
PCA-Hand (70)	44.94	34.27	15.73	20.22
+ distortion ($R = 10$)	56.74	34.83	14.61	15.73
+ δ-penalty	32.58	24.16	14.61	**14.04**

Table 2: System combination results using ROVER

System	DEL	INS	SUB	errors	WER %
system 1	10	3	12	25	14.04
system 2	9	4	13	26	14.61
system 3	11	4	17	32	17.98
system 4	5	4	16	25	14.04
ROVER	15	0	8	23	**12.9**

and/or virtual training samples are shown in Table 1. It can be seen that the proposed VSA method strongly improves the results for the appearance-based frame features. The usage of additional training data by virtual training samples (VTS) does not always lead to improvements (especially for the PCA-Hand features without distortion). However, a combination of both methods (i.e., virtual training samples extracted from visually aligned speaker sequences) leads to a WER of 14.04%, which is the best result reported for this data in the literature so far (17.98% in [3]).

Using the word confidences of the recognizer output, multiple recognition systems can easily combined by "rovering" over the system outputs [4]. We combined 4 different systems:

- system 1: sliding window over PCA-Frames
- system 2: sliding window over PCA-Frames and hand trajectory (HT) features
- system 3: sliding window over PCA-Frames and hand velocity (HV) features
- system 4: appearance-based PCA-hand patches

The results for the ROVER-based system combination are shown in Table 2. It can clearly be seen that the four systems, accounting well for different problems in sign language recognition (long words, short words, finger spelling, etc.), produce different word errors, and that a combination of the different systems leads to an improvement over the individual systems.

5 Conclusion

We presented a tracking adaptation method to obtain an adapted hand tracking path with optimized tracking positions w.r.t. recognition instead of a tracking criterion.

More robust models were trained using visual speaker alignments (VSA), to obtain speaker adapted features, and virtual training samples (VTS) easing the lack of data problem in vision based sign language recognition. The VSA and VTS adapted data improved the system performance in many cases, and the proposed method can be applied to any vision based system. In combination with a tracking path adaptation, the baseline WER of 44.94% on the benchmark database was improved to 14.04% WER, which is the currently best known WER for a single system on the used database. Furthermore, a ROVER-based system combination could improve the WER to 12.9%, which is the overall best known WER on the used database.

Interesting will be an iterative recognition and retraining of the system using the model adapted tracking path, and an analysis and extension of the proposed feature extraction methods.

References

[1] Chris J. C. Burges and B. Schölkopf. Improving the accuracy and speed of support vector machines. In *NIPS 97*, volume 9, pages 375–385, Vancouver, Canada, dec 1997.

[2] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. Tracking using dynamic programming for appearance-based sign language recognition. In *IEEE Automatic Face and Gesture Recognition*, pages 293–298, Southampton, April 2006.

[3] Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. Speech recognition techniques for a sign language recognition system. In *Interspeech 2007*, pages 2513–2516, Antwerp, Belgium, August 2007.

[4] J. Fiscus. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover). In *IEEE ASRU*, pages 347–352, Santa Barbara, CA, 1997.

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

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

[7] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[8] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Computer Vision & Image Understanding*, 81(3):358–384, March 2001.

[9] U. R. Wrobel. Referenz in Gebärdensprachen: Raum und Person. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, 37:25–50, 2001.

[10] Ruiduo Yang, Sudeep Sarkar, and Barbara Loeding. Enhanced level building algorithm to the movement epenthesis problem in sign language. In *CVPR*, MN, USA, June 2007.