Technology and Disability 20 (2008) 121–133
IOS Press

Spoken language processing techniques for sign language recognition and translation

Philippe Dreuw*, Daniel Stein, Thomas Deselaers, David Rybach, Morteza Zahedi, Jan Bungeroth and Hermann Ney
Human Language Technology and Pattern Recognition, Computer Science Department 6, RWTH Aachen University, Germany

Abstract. We present an approach to automatically recognize sign language and translate it into a spoken language. A system to address these tasks is created based on state-of-the-art techniques from statistical machine translation, speech recognition, and image processing research. Such a system is necessary for communication between deaf and hearing people. The communication is otherwise nearly impossible due to missing sign language skills on the hearing side, and the low reading and writing skills on the deaf side. As opposed to most current approaches, which focus on the recognition of isolated signs only, we present a system that recognizes complete sentences in sign language. Similar to speech recognition, we have to deal with temporal sequences. Instead of the acoustic signal in speech recognition, we process a video signal as input. Therefore, we use a speech recognition system to obtain a textual representation of the signed sentences. This intermediate representation is then fed into a statistical machine translation system to create a translation into a spoken language. To achieve good results, some particularities of sign languages are considered in both systems. We use a publicly available corpus to show the performance of the proposed system and report very promising results.

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. Although deaf, hard of hearing and hearing signers can fully communicate among themselves by sign language, there is a big communication barrier between signers and hearing people without signing skills. Here, we propose a sign-to-speech communication system to aid the signing community with their everyday communication problems. Figure 1 illustrates the various components necessary for such a system.

Linguistic research in sign language has shown that signs mainly consist of four basic manual components [20]: hand configuration, place of articulation, hand movement, and hand orientation. Additionally, non-manual components like facial expression and body posture are used.

In [16,26] reviews on recent research in sign language and gesture recognition are presented. In vision-based *automatic sign language recognition* (*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 [28], and are person dependent [2, 23]. Furthermore, most approaches focus on the recognition of isolated signs only [2,23], or on the simpler case of gesture recognition [24] for small vocabularies. 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.

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. As we point out in Section 2,

*Corresponding author. E-mail: dreuw@cs.rwth-aachen.de.

ISSN 1055-4181/08/$17.00

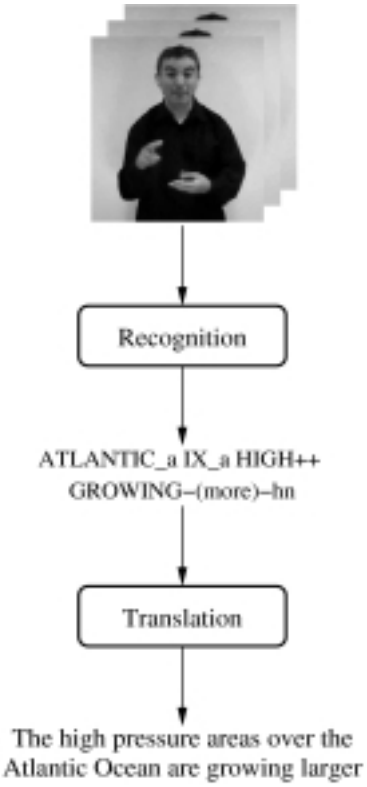

Fig. 1. Complete system setup with an example sentence: After automatically recognizing the input sign language video, the translation module has to convert the intermediate text format (glosses) into written text.

the recognition part of our work is based on a large vocabulary speech recognition system [13]. In particular, we present a complete vision-based framework for person independent continuous sign language recognition opposed to isolated gesture-recognition works presented by most other authors [2,23,24], and analyze the impacts of common speech recognition techniques in sign language recognition on a publicly available database with several speakers.

As mentioned above, recognition is only the first step of a sign-language to spoken-language system. The intermediate representation of the recognized signs is further processed in an automatic machine translation system to create a spoken language translation, as discussed in Section 3. The machine translation system accounts for the different grammar and vocabulary of the sign language. To enhance translation quality, we also propose to use visual features from the recognition process and include them into the translation as an additional knowledge source.

In Section 4, we present our experimental setup with some promising results. Finally, we give a conclusion of the experimental results and an outlook in Section 5.

2. Speech and sign language recognition

Automatic speech recognition (*ASR*) is the conversion of an acoustic signal (sound) into a sequence of written words (text). Related tasks to speech recognition are e.g.:

- Speech understanding: generating a semantic representation
- Speaker recognition: identifying the person who spoke
- Speech detection: separating speech from non-speech

On the signal level, further tasks are e.g.:

- Speech enhancement: improving the intelligibility of a signal
- Speech compression: encoding speech signal for transmission or storage with a small number of bits.

Due to the high variability of the speech signal, speech recognition – outside lab conditions – is known to be a hard problem. Most decisions in speech recognition are interdependent, as word and phoneme boundaries are not visible in the acoustic signal, and the speaking rate varies. Therefore, decisions cannot be drawn independently but have to be made within a certain context, leading to systems that recognize whole sentences rather than single words.

One of the key idea in speech recognition is to put all ambiguities into probability distributions (so called stochastic knowledge sources, see Fig. 2). Then, by a stochastic modelling of the phoneme and word models, a pronunciation lexicon and a language model, the free parameters of the speech recognition framework are optimized using a large training data set. Finally, all the interdependencies and ambiguities are considered jointly in a search process which tries to find the best textual representation of the captured audio signal. In contrast, rule-based approaches try to solve the problems more or less independently.

In order to design a speech recognition system, four crucial problems have to be solved:

1. preprocessing and feature extraction of the input signal,

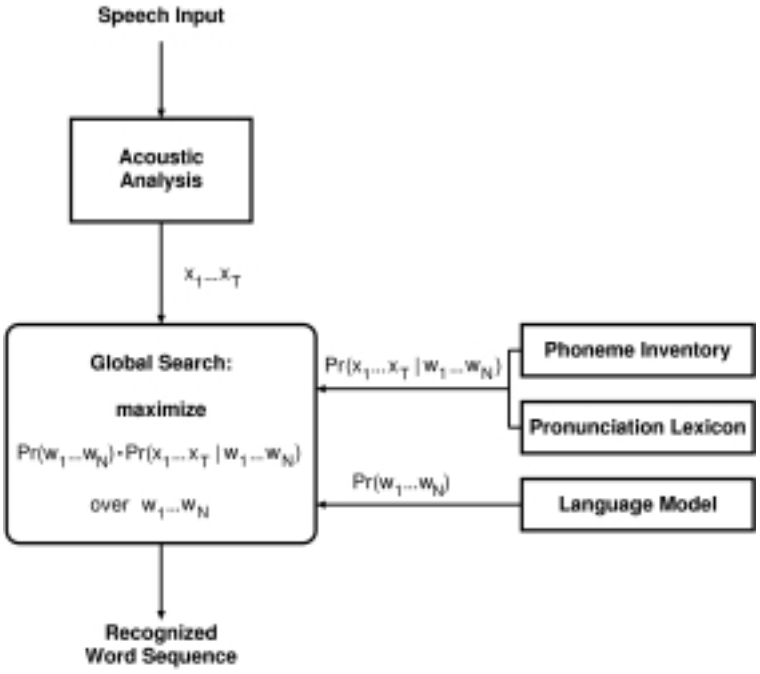

Fig. 2. Bayes' decision rule used in speech recognition. For a given audio input sequence, features are extracted to be used in a global search of the models, i.e. a word sequence, which best describe the current observation.

2. specification of models and structures for the words to be recognized,
3. learning of the free model parameters from the training data, and
4. search the maximum probability over all models during recognition (see Fig. 2).

2.1. Sign language recognition

We call the conversion of a video signal (images) into a sequence of written words (text) *automatic sign language recognition* (*ASLR*). We propose to use the knowledge obtained in speech recognition research over the last decades to create a sign language recognition system. In particular, we use a state-of-the-art large vocabulary speech recognition system as a basis [13], since the similarities between both tasks are great: Similar to spoken languages we have to process temporal sequences of input data. However, in sign language recognition we have to deal with visual observations instead of acoustic observations.

In order to build a robust recognition system which can recognize continuous sign language speaker independently, we have to cope with various difficulties:

1. *coarticulation:* the appearance of a sign depends on the preceding and succeeding signs.
2. *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.

2.2. Problems and differences in comparison to ASR

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.

2.2.1. Simultaneousness

One major issue in sign language recognition compared to speech recognition is the possible simultaneousness: a signer can use different communication channels (facial expression, hand movement, and body posture) in parallel. For example, different comparative degrees of adjectives are indicated through increased facial expression, indirect speech through spatial geometry of the upper part of the body, noun-to-verb derivation through increased speed and reduction

of the signing space; all this happens while the subject is still signing normally.

2.2.2. 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 [25]. A challenging task is to define a model for spatial information containing the entities created during the sign language discourse. An example for the use of virtual signing might be the simple looking sentence "*he* gives *her* a book": Such a sentence would cause (under normal circumstances) no problems to modern ASR frameworks. However, it would be quite a complex problem in sign language recognition, as one would have to use context knowledge in order to know where the "male" and "female" persons are located in the virtual signing space (see also Section 3.2).

2.2.3. 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 (see Fig. 3).

2.2.4. 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.

2.2.5. 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 [23,27]. Movement epenthesis refers to movements which occur regularly in natural sign language in order to change the location in signing space. Movement epenthesis conveys no meaning in itself but rather changes the meaning of succeeding signs, e.g. to express that the wind is blowing from north-to-south instead of south-to-north.

2.2.6. 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.

2.2.7. Whole-word models and sub-word units

The use of whole-word models for the recognition of sign language with a large vocabulary is unsuitable, as there is usually not enough training material available to robustly train the parameters of the individual word models. According to the *linguistic* work on sign language by Stokoe [20], a phonological model for sign language can be defined, dividing signs into units. In ASR, words are modelled as a concatenated sub-word units. These sub-word units are shared among the different word-models and thus the available training material is distributed over all word-models. On the one hand, this leads to better statistical models for the sub-word units, and on the other hand it allows to recognize words which have never been seen in the training procedure. For sign language *recognition*, however, no suitable decomposition of words into sub-word units is currently known.

One of the challenges in the recognition of continuous sign language on large corpora is the definition and modelling of the basic building blocks of sign language. These sub-word units are similar to phonemes in ASR. Inspired by linguistic research, the signs could be broken down into their constituent visemes, such as the hand shapes, types of hand movements, and body locations at which signs are executed. Furthermore, they will allow the consideration of context dependency with new suitable models for within-word coarticulation (e.g. diphones or triphones).

2.3. System overview and feature description

Our ASLR system is based on Bayes' decision rule: the word sequence which best explains the current observation given the learned model is the recognition result. For data capturing we use standard video cameras rather than special data acquisition devices. To

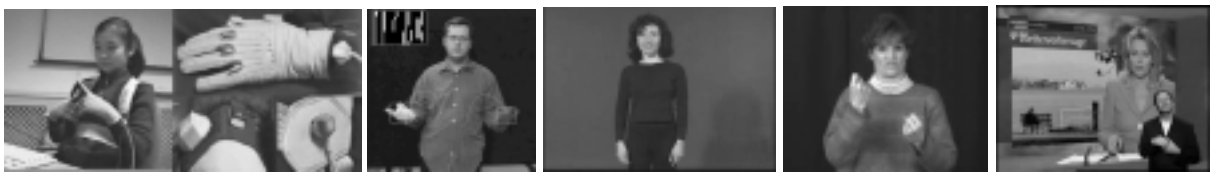

Fig. 3. Different environment assumptions resulting in completely different problems in feature extraction (f.l.t.r.): data gloves, colored gloves, blue-boxing, unconstrained with static background, and unconstrained with moving and cluttered background.

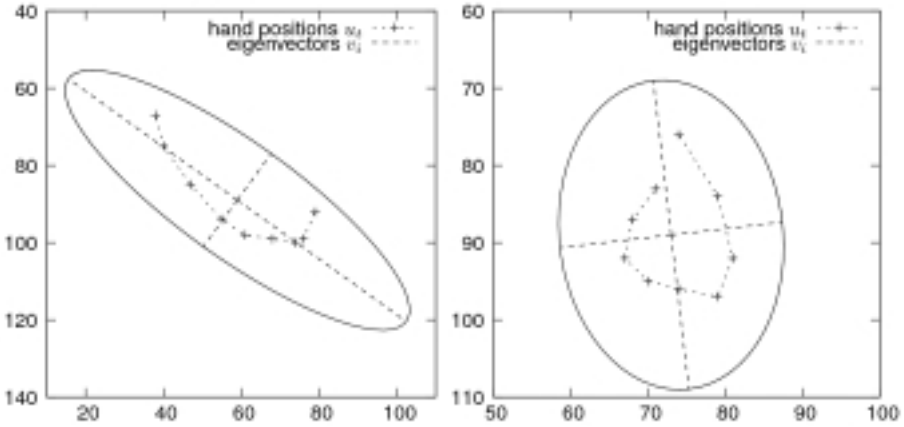

Fig. 4. Examples of different hand trajectories and corresponding eigenvectors for $\delta = 4$. The covariance matrices are visualized as ellipses with axes of length $\sqrt{\lambda_i}$.

model the video signal we use appearance-based features. To cope with the many difficulties described above, new models have to be developed that are more robust against noise, to the visual appearance of the signers, and to accentuate speech (such as male and female speakers) covering more languages.

2.3.1. Visual modeling

As it is still unclear how sign language words can be split up into sub-word units, e.g. phonemes, suitable for sign language recognition, our corpus (c.f. Section 4.1) is annotated in *glosses*, i.e. whole-word transcriptions (see Section 3.2), and the system is based on whole-word models. This means for Figure 2 that the phoneme inventory in combination with a pronunciation lexicon is replaced by a word model inventory without a lexicon. Each word model consists of several *pseudo-phonemes* modeling the average word length seen in training. Each such phoneme is modeled by a 3-state left-to-right hidden Markov model (HMM) with three separate Gaussian mixture models (GMM) and a globally pooled diagonal covariance matrix [8].

Due to various dialects in natural 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 Swiss sign language [3]). Small differences between the appearance and the length of the utterances are compensated by the HMMs, but different pronunciations of a sign must be modelled by separate models, i.e. a different number of states and different GMMs. Therefore, we added pronunciation information to the corpus annotations and adjusted our language models.

2.3.2. Language models

The language models aim at representing syntax and semantics of natural language (spoken or written). They are needed in automatic language processing systems that process speech (i.e. spoken language) or language (i.e. written language).

In our first approach, where all communication channels in sign language are considered at once, the grammatical differences to a spoken language do not pose problems in the recognition framework. They are modeled by statistical language models as in ASR. Language models based on the sign level versus independent language models for each communication channel (e.g. the hands, the face, or the body) could be analyzed, too. The former means an early integration of the features, the latter a late fusion of the systems (also compare Fig. 5 with Fig. 6).

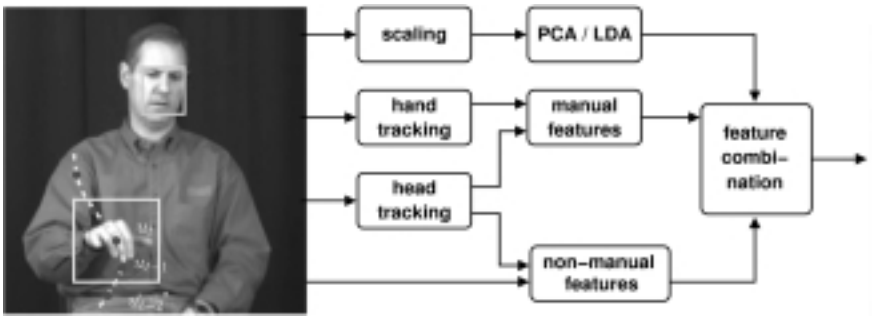

Fig. 5. Feature combination on the signal level.

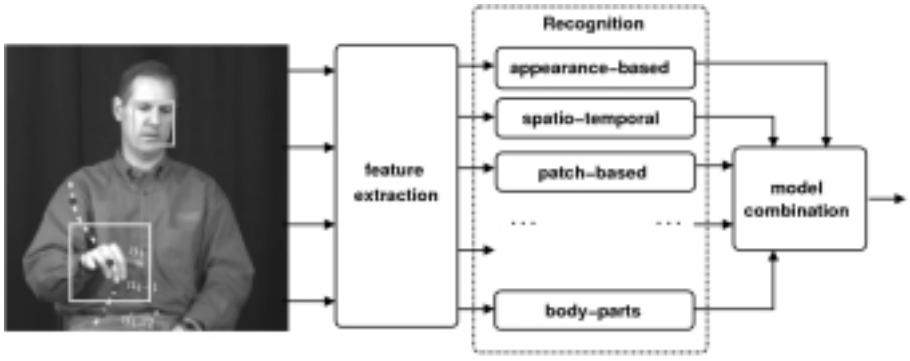

Fig. 6. Combination of different approaches to recognize sign language on the model level.

In Bayes' decision rule (see Fig. 2), 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:

$$\text{argmax}_{w_1^N} \left\{ Pr^{\alpha}(w_1^N) \cdot Pr^{\beta}(x_1^T|w_1^N) \right\} =$$

$$\text{argmax}_{w_1^N} \left\{ \frac{\alpha}{\beta} \cdot \log Pr(w_1^N) + \log Pr(x_1^T|w_1^N) \right\}$$

The factor α/β is referred to as *language model factor*. The LM was generated on the training sentences using the SRILM toolkit [21].

2.3.3. Appearance-based features

Many research groups in gesture or sign language recognition use quite complex methods to recognize the gestures, like fingertip detection, calculating the angles between the fingers or matching of 3D-models. Often used features for vision-based gesture and sign language recognition are

- color: brightness, skin color models, etc.
- texture: Gabor-filters, gradients, etc.
- shape: active shapes, active contour models, etc.
- motion: centroids, difference images, optical flow, etc.

Spatio-temporal segmentation of video sequences is also an often used and essential step in video analysis. It attempts to extract backgrounds and independent objects in the dynamic scenes captured in the sequences. In an appearance-based approach one does not create such modular systems which have to extract specific features from different body parts, even though all the information one needs to recognize a gesture is encoded in the image itself – segmenting images is very difficult and never perfect.

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 [6]. They give a global description of all (manual and non-manual) features proposed in linguistic research. In subsequent steps, this baseline feature is extended by features accounting for the hands and their positions.

2.3.4. Manual features

Glove-based systems offer the immediate extraction of manual features while hindering a natural way of signing. In our vision-based approach, 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 the image sequences: 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 [7]. 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.

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 over the observed hand positions during that time period [8]. 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 fit 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 4 shows some examples of trajectories and their eigenvectors and eigenvalues. The hand trajectory (HT) features presented here are similar to the features presented in [23].

2.3.5. Feature selection and combination

In [10] it has been shown that the combination of different models or features leads to an improvement over the individual models. By a linear combination of features describing different characteristic parts of the language, i.e. vector components of different knowledge sources, and a suitable scaling and weighting of features, the quality of the recognizer can be strongly improved.

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 *Principal Component Analysis* (*PCA*) or *Linear Discriminant Analysis* (*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, successive 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 5 shows how we extract and combine features.

Another type of feature combination can be done on the model level (see Fig. 6) by a log-linear combination of independently trained models. For example, a spatio-temporal based hand model accounting for hand motion, and an appearance-based hand model accounting for the different hand shapes. The model weights have to be 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. The results achieved using different features and combination methods are presented in Section 4.

A third possible combination of the features extracted from the different communication channels could be done by a combination on the system level. The late fusion of specialized systems for independent communication channels could be analyzed, e.g. by recogniser output voting error reduction (ROVER) [9], for decision fusion of concurring systems for the same or different communication channels in sign language.

3. Machine translation

Automatic machine translation is the translation from a source language into a target language by means of either data-based or rule-based methods. For rule-based systems, a set of translation rules has to be created manually by bilingual language experts, while for data-based approaches the machine has to derive the rules itself by extracting them from given examples (*supervised learning*), without any prior language or grammar knowledge involved.

Statistical machine translation (SMT) is a data-based translation method that was initially inspired by the so-called noisy-channel approach: the source language is interpreted as an encryption of the target language, and thus the translation algorithm is typically called a decoder. In practice, statistical machine translation often outperforms rule-based translation significantly on international translation challenges, given a sufficient amount of training data.

One of the ground-breaking papers in this area was based on experiments from French to English, thus the

source language is usually denoted f and the target language e [4]. We train several models based on a given bilingual collection of previously translated sentences, known as the corpus (e.g. parliament speeches, bible, etc.) The first thing to learn from a corpus is a mapping of corresponding words, the *alignment*. Afterwards, to translate a given sentence consisting of J words $f_1 \dots f_J$, we create all possible target sentences $e_1^I = e_1 \dots e_I$ and assign a probability to them based on the experiences we made in the training. The sentence that maximizes the a-posteriori probability $Pr(e_1^I|f_1^J)$, is then selected as the best translation. While the initial approach was based on the Bayes' decision rule as in ASR, nowadays state-of-the-art decoder typically employ a log-linear feature model that combines several models h_m with scaling factors λ_m:

$$p(e_1^I|f_1^J) = \frac{1}{Z(f_1^J)} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right). \quad (1)$$

We can ignore the denominator function Z in the actual translation since it only normalizes the probability distribution and does not change the maximizing sentence.

An example for an additional knowledge source is the language model that prefers valid sentences in the target language over invalid sentences, penalizing for example "There hello" compared to "Hello there".

The scaling factors in Eq. (1) are commonly learned automatically from training data using stochastic optimization techniques.

3.1. Sign language translation

Similar to the sign language recognition based on speech recognition, we propose to also use the methods derived for spoken language translation. Since we are using a written form, the adaptation does not seem to be very hard at first, but there are some striking differences for sign languages that pose several challenges explained below.

While the first papers on sign language translations only date back to roughly a decade [22] and typically employed rule-based systems, several research groups have recently focussed on data-driven approaches. In [18], a SMT system has been developed for German and German sign language in the domain weather reports. Their work describes the addition of pre- and post-processing steps to improve the translation for this language pairing. The authors of [14] have explored example-based MT approaches for the language pair English and sign language of the Netherlands with further developments being made in the area of Irish sign language. In [5], a system is presented for the language pair Chinese and Taiwanese sign language. The optimizing methodologies are shown to outperform a simple SMT model. In the work of [17], some basic research is done on Spanish and Spanish sign language with a focus on a speech to gesture architecture.

3.2. Problems and differences in comparison to MT

Apart from being poorly resourced languages, sign languages also pose specific problems for MT due to their modality:

3.2.1. Annotation

One of the biggest obstacles is that for sign languages no official written form exists. If the transcription and accordingly the recognition output is too complex, the data sparseness prevents any meaningful training of the models. This means that for the notation system used in this process, a good trade-off between accuracy and generality has to be found that serves as a useful intermediate step.

The first attempts for a notation system for sign languages are made in the 1960's by Stokoe [20]. Stokoe argued that there are three aspects of manual sign articulation, namely hand configuration, place of articulation and movement. This model was extended by the hand situation as a fourth parameter [1,12]. For our purpose, however, they rely too heavily on the syntactic components, while we are more interested in the semantic meaning.

We use glosses as a semantic representation of the sign language. As a convention, the meaning of the sign is written as the upper case stem form of the corresponding word in a spoken language [15]. For our translation, it annotates all important sign language grammar features, some of which to be mentioned below:

3.2.2. Word flexion

Most known sign languages belong to the group of languages where the word flexion is more important than the word position in the sentence. Flexed verbs usually share the same root, which means that they are mostly identical in their components, but differ in such elements as movement speed, direction or amount of signing space used. The direction of a verb indicates subject and object and number of occurrences, using a predefined set of movements to differ between casus and numerus. As seen in [18], an automatic grammar

parser for the source language can be used as an external knowledge source to improve the translation quality, especially for small corpora, but for these phenomenons in sign language, no parser exists.

3.2.3. Discourse entities

Entities like persons or objects can be stored in the sign language space by executing them at a certain location and later just referencing them by pointing to the space [25]. By flexing a verb towards this so-called discourse entity, the signer is referring back to this person like in a pronoun ("She is giving him an apple"). Normally, the starting point is referencing to the subject and the end point to the direct or indirect object. This technique is called *verbal agreement*. Since every location is clearly defined in regards of what person it references, verbal agreement is in some ways more exact than verbal flexion in some vocal languages. Locations can also be used to reference objects, abstract concepts or sentences. For translation, this means that the recognized sign must contain all the details given by the deaf person in order not to confuse their relationship amongst each other.

We will give an example in gloss annotation that captures some of the phenomenons mentioned above. The gloss sentence "ATLANTIC_a IX_a HIGH++ GROWING (more)-hn" can be translated into English with "The high pressure areas over the Atlantic Ocean are growing larger" . The three signs are transcribed with the glosses "HIGH", "ATLANTIC" and "GROWING" representing their meaning in English. The sign "IX" is a pointing gesture to reference the same space "_a" used by the discourse entity "ATLANTIC". Signs repeated (for example to indicate plural forms) are annotated with a double-plus, mouth pictures are written in brackets, e.g. "(more)", "-hn" means that the signer is nodding during signing. The corresponding alignment for this sentence can be seen in Fig. 7.

In this example, it also becomes apparent that the word order is different in the languages so that we have to think about reordering during the decoding step. For translation, admitting all possible permutations would be computationally too expensive for large sentences, so we usually allow only a limited set of permutations, translate all as if they were normal input sentences, and choose the most probable translation out of them.

3.3. Sign-to-speech

Speech to speech systems that combine spoken language recognition with translation systems already exist and work reasonably well. Some attempts on sign-to-speech have already been made [19].

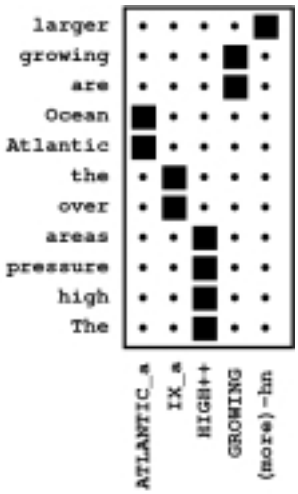

Fig. 7. Alignment for English and ASL. Squares represent automatically derived word mappings.

But how should we handle the morpho-syntactic complexeties in poorly resourced sign language data collections? In this work, we recognize the "stemmed" gloss, that is, the gloss that does not contain all the spatial information but just indicates that the signer is executing this particular sign somewhere in front of his body. As to what it references and to where it was pointed at, we use the visual features derived during the recognition process (e.g. tracked hand positions) as a kind of additional part-of-speech and flexion information and pass it on to the decoder. For example, in the recognition system of [8], the position of the signing hand is localized automatically through movement derivation, and we can tell the decoder that it has seen a pointing finger (deixis) with a gloss. The tracking information already contains the necessary information needed to differentiate between a single reference or a location sign (e.g. "this woman" vs. "a woman over there"). In preliminary works this helped to improve the overall performance, as presented in the experiment section below.

4. Experimental results

To benchmark our system, we use the publicly available RWTH-Boston-104 corpus presented in [8].

Table 1
RWTH-Boston-104 corpus statistics

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

Table 2
Baseline results

Appearance-based features	Dim.	WER [%]
intensity (w/o pron.)	1024	54.0
intensity (w/ pron.)	1024	37.0
intensity (w/ pron. + tangent dist.)	1024	33.7
motion (pixel based)	1024	51.1
intensity+motion	2048	42.1

Table 3
Results for feature combinations with hand features

Features	Dim.	[% WER]
PCA-frame	110	27.5
+ hand-position (HP)	112	25.3
+ hand-velocity (HV)	112	24.2
+ hand-trajectory (HT)	112	23.6
model-combination [8]	2×100	17.9

4.1. RWTH-Boston-104 Database

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 [15]. 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, see Fig. 8) 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 (e.g. words seen once in training). The sentences have a rather simple structure. The test corpus has one out-of-vocabulary (OOV) word which cannot be recognized correctly using whole-word models.

4.2. Results

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.

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

We use only unseen data from the test sentences for evaluation. A common performance measure is the word error rate (WER) representing the minimum number of substitution, deletion and insertion errors divided by the total number of signs in the recognized sentence.

4.2.1. Baseline

First, we analyze different appearance-based features for our baseline system. The baseline system is Viterbi trained and uses a trigram LM (c.f. Section 2.3.2). 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 a distance measure accounting for global image transformations [6] 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 most simple 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 work reasonably well in previous works [6]. Another important point is the usage of pronunciation modelling in sign language: it can be seen that by adding pronunciation information to the corpus and the adjustment of the used trigram language model, the system performance can already be improved from 54.0% to 37.0% WER.

4.2.2. Feature reduction

Obviously, the high dimensional appearance-based feature vectors include a lot of background (noise) and one would need many more observations to train a robust model. To reduce the feature dimension and to eliminate background 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 theoretically LDA should be better suited for pattern recognition tasks, here the training data is insufficient for a numerically stable estimation of the LDA transformation. PCA, which is reported to be more stable for high dimensional data with small training sets, outperforms LDA.

Fig. 8. Some examples of the RWTH-Boston-104 database showing the 3 different speakers.

4.2.3. 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 Figs 9 and 10 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. Furthermore, windowing with large temporal contexts increases the system performance, as coarticulation effects are described now.

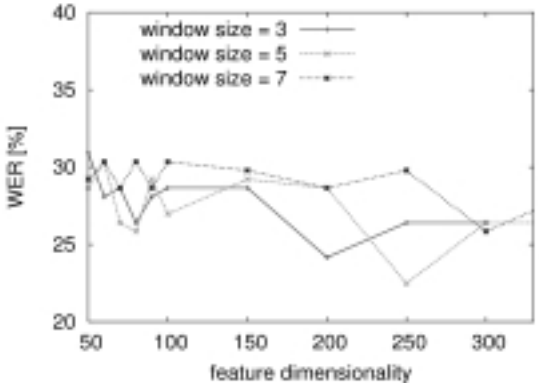

Fig. 9. Combination of PCA-frames using LDA windowing.

4.2.4. 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 (c.f. Section 2.3.5). 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, 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. 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. On the contrary, 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 disturbed for short words (i.e. less than 5 states) because at the word boundaries strong coarticulation effects occur.

4.2.5. Language model

Figure 11 shows the effect of using different -gram language models and scales. As in ASR, the usage of language models in combination with the added sign language pronunciation information achieve large improvements (c.f. baseline results). Interestingly, the achieved improvement factors are similar to those from speech recognition [11]. 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.

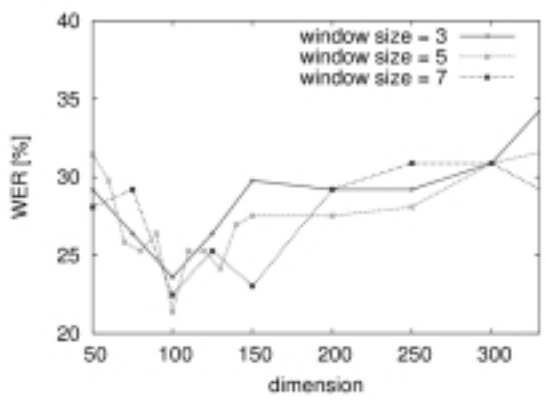

Fig. 10. Combination of PCA-frames using PCA windowing.

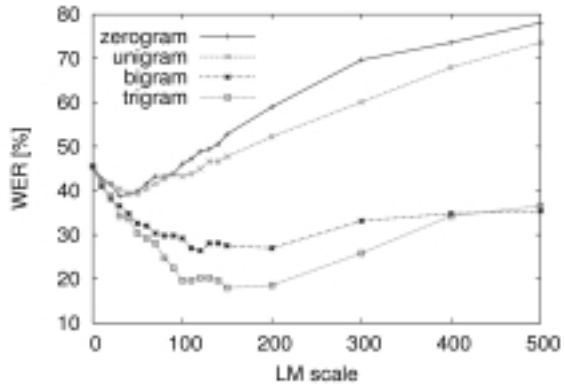

Fig. 11. Results for different LMs and scales.

4.2.6. Sign-to-speech translation

On the best recognition result, we achieve an overall system performance of a signed-video-to-written-English translation of WER, which is a very reasonable quality and, in spite of glosses, is intelligible for most people. In another set of experiments, for incorporation of the tracking data, the tracking positions of the dominant-hand were clustered and their mean calculated. Then, for deictic signs, the nearest cluster according to the Euclidean distance was added as additional word information for the translation model. For example, the sentence JOHN GIVE WOMAN IX COAT might be translated into *John gives the woman the coat* or *John gives the woman over there the coat* depending on the nature of the pointing gesture IX. This helped the translation system to discriminate between deixis as distinctive article, locative or discourse entity reference function in preliminary test runs to reduce the error rates by 2%.

5. Summary and conclusion

We presented a system that can automatically recognize sign language and translate it into a spoken language aiming at reducing the communication barrier for deaf and hard-of-hearing people. The system is composed of two main components: (1) a sign-language recognition part that has a video signal as input and creates an intermediate text representation of the signed gestures and (2) a translation part which creates a spoken language translation from the intermediate textual representation of the signs.

The recognition part is based on recent developments in automatic speech recognition and image processing. We have shown that many of the principles known from ASR, such as pronunciation and language modelling, can be transferred to the new domain of vision-based continuous ASLR. In particular, we have shown that appearance-based features are well suited for the recognition of sign-language and that therefore special data acquisition tools are not necessary.

We present very promising results on a publicly available benchmark database of several speakers consisting of videos. It is shown that many of the difficulties that occur in ASLR can be addressed by the means of appropriate feature extraction techniques and a suitable selection of visual features. The translation part is based on a modern statistical machine translation system where the preprocessing stage was custom-built for the language pair ASL/English. In informal experiments on other sign language/spoken language pairs, similar methods perform equally well.

To the best of our knowledge, the presented approach is the first one to combine data-driven methods for the recognition and the translation of sign languages into spoken languages. However, only the combination of these methods can lead to systems that bring huge improvements in communication for deaf people. With the system proposed, we are able to produce a unique sign-language-to-speech system at a reasonable quality for small to medium vocabulary sizes.

5.1. Outlook

In many of the fields touched in this paper, important tasks are still unsolved. We expect large improvements in the recognition phase of the system by incorporating a better model of the human body configuration. A suitable definition of sub-word units would also be very helpful and it would probably also alleviate the burden of insufficient data for model creation.

For the translation step, preliminary experiments have shown that the incorporation of the tracking data for deixis words helps to properly interpret the meaning of the deictic gestures. Other features that are likely to improve the error rates include velocity movements, tilt of the head, and shifts of the upper body. Furthermore, a thorough analysis of the entities used in a discourse is required to properly handle pronouns.

References

[1] R. Battison, *Lexical Borrowing in American Sign Language*, Linstok Press, MD, USA, 1978.

[2] R. Bowden, D. Windridge, T. Kadir, A. Zisserman and M. Brady, A linguistic feature vector for the visual interpretation of sign language, in: *ECCV*, (vol. 1), 2004, pp. 390–401.

[3] Penny Boyes Braem, *Einführung in die Gebärdensprache und ihre Erforschung*, Signum-Verlag, 1995.

[4] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra and R.L. Mercer, The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics* **19**(2) (June 1993), 263–311.

[5] Y.-H. Chiu, C.-H. Wu, H.-Y. Su and C.-J. Cheng, Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis, *IEEE Trans PAMI* **29**(1) (2007), 28–39.

[6] P. Dreuw, T. Deselaers, D. Keysers and H. Ney, Modeling image variability in appearance-based gesture recognition, in: *Statistical Methods in Multi-Image and Video Processing*, Graz, Austria, May 2006, pp. 7–18.

[7] 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*, Southampton, April 2006, pp. 293–298.

[8] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi and H. Ney, Speech recognition techniques for a sign language recognition system, in *ICSLP*, Antwerp, Belgium, August 2007.

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

[10] J. Kittler, On combining classifiers, *IEEE Trans PAMI* **20**(3) (March 1998), 226–239.

[11] D. Klakow and J. Peters, Testing the correlation of word error rate and perplexity, *Speech Communication* **38** (2002), 19–28.

[12] E.S. Klima and U. Bellugi, *The Signs of Language*, Harvard University Press, Cambridge, UK, 1979.

[13] 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, September 2006.

[14] S. Morrissey and A. Way, An Example-based Approach to Translating Sign Language, in: *Workshop in Example-Based Machine Translation* (*MT Summit X*), Phuket, Thailand, 2005, pp. 109–116.

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

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

[17] R. San-Segundo, R. Barra, L.F. D'Haro, J.M. Montero, R. Córdoba and J. Ferreiros, A Spanish Speech to Sign Language Translation System for assisting deaf-mute people, in: *ICSLP*, Pittsburgh, PA, 2006.

[18] D. Stein, J. Bungeroth and H. Ney, Morpho-Syntax Based Statistical Methods for Sign Language Translation, in: *11th EAMT*, Oslo, Norway, June 2006, pp. 169–177.

[19] D. Stein, P. Dreuw, H. Ney, S. Morrissey and A. Way, Hand in Hand: Automatic Sign Language to Speech Translation, in: *The 11th Conference on Theoretical and Methodological Issues in Machine Translation*, Skoevde, Sweden, September 2007.

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

[21] A. Stolcke, SRILM – an extensible language modeling toolkit, in: *ICSLP*, (vol 2), Denver, CO, September 2002, pp. 901–904.

[22] T. Veale, A. Conway and B. Collins, The Challenges of Cross-Modal Translation: English to Sign Language Translation in the ZARDOZ System, *Journal of Machine Translation* **13**(1) (1998), 81–106.

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

[24] S.B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian and T. Darrell, Hidden conditional random fields for gesture recognition, in: *CVPR*, (vol. 2), New York, USA, June 2006, pp. 1521–1527.

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

[26] T.S. Huang and Y. Wu, Vision-based gesture recognition: a review, in: *Gesture Workshop*, (vol. 1739) of *LNCS*, Gif-sur-Yvette, France, March 1999, pp. 103–115.

[27] 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.

[28] G. Yao, H. Yao, X. Liu and F. Jiang, Real time large vocabulary continuous sign language recognition based on op/viterbi algorithm, in: *ICPR*, (vol. 3), Hong Kong, August 2006, pp. 312–315.