Towards an Automatic Sign Language Translation System

Britta Bauer*, Sonja Nießen‡ and Hermann Hienz*

*Department of Technical Computer Science
Aachen University of Technology (RWTH)
D-52056 Aachen, Germany
http://www.techinfo.rwth-aachen.de/
{bauer,hienz}@techinfo.rwth-aachen.de

‡Department 6 of Computer Science
Aachen University of Technology (RWTH)
D-52056 Aachen, Germany
http://www.informatik.rwth-aachen.de/I6/
niessen@informatik.rwth-aachen.de

Abstract

This paper outlines a concept of a continuous sign language translation system. The system consists of two main components. One of these is a video-based recognition tool, which aims for automatic signer dependent recognition of sign language sentences, based on a lexicon of 100 signs of German Sign Language. A single colour video camera is used for image recording. The recognition is based on Hidden Markov Models concentrating on manual sign parameters. The video-based recognition tool achieves an accuracy of 91.6%. The second main component is a translation tool, that translates from sequences of sign identifiers (provided by the recognition tool) into text. It uses two types of statistical information sources: a translation model and a language model. The translation model is decomposed into a lexical model and an alignment model. Experiments on tasks with comparable complexity yeald accuracies between 88 and 100%.

Keywords: Sign language translation system, sign language recognition, statistical pattern recognition, Hidden Markov Models, gesture recognition, statistical machine translation.

1 Introduction

Sign languages, although different in form, serve the same functions as a spoken language. They are natural languages which are used by many deaf people all over the world, e.g., GSL (German Sign Language) in Germany or ASL (American Sign Language) in the United States. As a minority community, with limited access to heard speech, deaf people rely on the relatively small numbers of hearing people who learn sign language, to be their interpreters. Since sign languages have different grammatical structures to European spoken languages translation is a problem in most contexts without provision of highly trained interpreters. Yet advances in linguistic research on sign languages and in new statistical methods in the field of machine translation offer models of the way in which the languages work and offer for the first time, the possibility that automated language translation as applied to spoken and written languages may be applied to sign language. The development of a system for translating sign language into spoken language would be of great help for deaf as well as hearing people.

This paper outlines a system design for an automatic translation system sign-to-text, that is composed of a video-based continuous sign language recognition tool and a statistical translation tool translating sequences of signs into natural text.

1.1 Related Work

This section describes a system which is concerned with the translation of japanese sign language. It is not an aim of this section to describe other sign language recognition systems, i.e. [7] [11] [13].

The dialogue system of [6] for Japanese Sign Language is based on a sign sentence recognition tool for 38 signs. Additional main modules of the system are: sign-language synthesis, and dialogue control. For the recognition part a stereo camera and a pair of colored gloves to track the three-dimensional movements of the signer is employed. The sentences are signed with additional pauses between all signs. The translation is achieved by a set of limited standard sentences, which are defined in a developed lexicon. For the sign-language synthesis signer motion data is used. A commercial optical motion capture system and a pair of data-gloves are used to obtain motion data. The dialogue control module makes all modules work together to accomplish a simple “ask-answer” dialogue. The treatment of occured errors is not possible.

1.2 Theory of Hidden Markov Models

This section briefly discusses the theory of Hidden Markov Models (HMMs). A more detailed description of this topic can be found in [9, 10].

Given a set of N states s_i we can describe the transitions from one state to another at each time step t as a stochastic process. Assuming that the state-transition probability a_{ij} from state s_i to state s_j only depends on preceding states, we call this process a Markov chain. The further assumption, that the actual transition only depends on the very preceding state leads to a first order Markov chain. We can now define a second stochastic process that produces at each time step t symbol vectors x. The output probability of a vector x only depends on the actual state, but not on the way the state was reached. The output probability density $b_i(x)$ for vector x at state s_i can either be discrete or continuous. This double stochastic process is called a HMM.

In our approach each sign is modelled with one HMM. Figure 1 illustrates the modelling of a sequence of continuous signs with one HMM for each sign.

The first row of the figure shows two images of the sign DAS, four images of the sign WIEVIEL and three images of the sign KOSTEN as recorded by the video camera. For each image a feature vector X is calculated. The sequence of feature vectors represents the observation sequence O. The order of visited states forms the state sequence. With Bakis topology for each HMM, the system is able to compensate different speeds of signing. An initial state of a sign can only be reached from the last state of a previous model.

1.3 The Statistical Approach to Translation

The goal is the translation of an input sentence given in some source language into a target language. We are given a source string $f_1^J = f_1 ... f_j ... f_J$, which is to be translated into a target string $e_1^I = e_1 ... e_i ... e_I$. Among all possible target strings, we will choose the string with the highest probability which is given by Bayes' decision rule [3]:

$$\left(I_{opt}, \hat{e}_1^{I_{opt}}\right) = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\} . \tag{1}$$

$Pr(e_1^I)$ is the language model (LM) of the target language, whereas $Pr(f_1^J | e_1^I)$ is the translation model. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language.

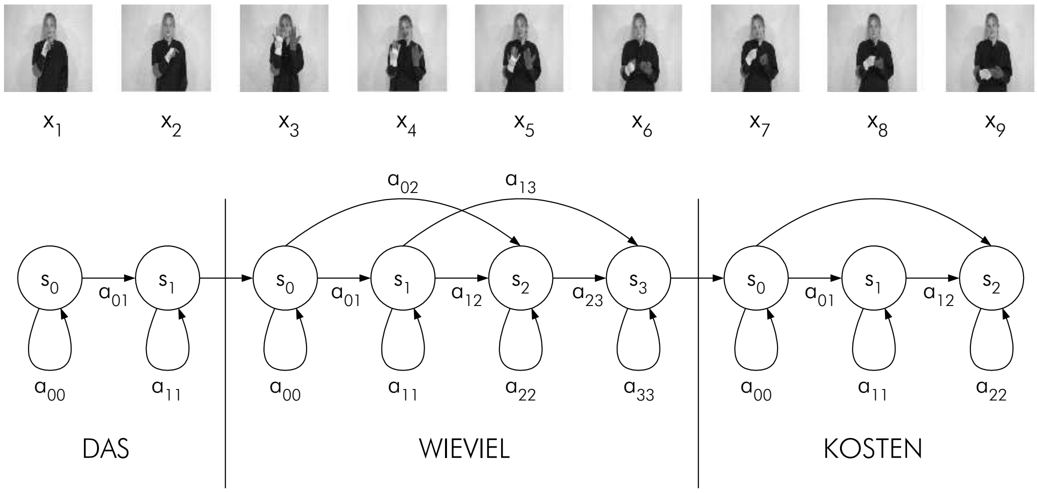

Figure 1: Image sequence of the sentence DAS WIEVIEL KOSTEN ('How much does that cost?'). Modelling each sign of the sentence with one Bakis-HMM.

The overall architecture of the statistical translation approach is summarized in Fig. 2.

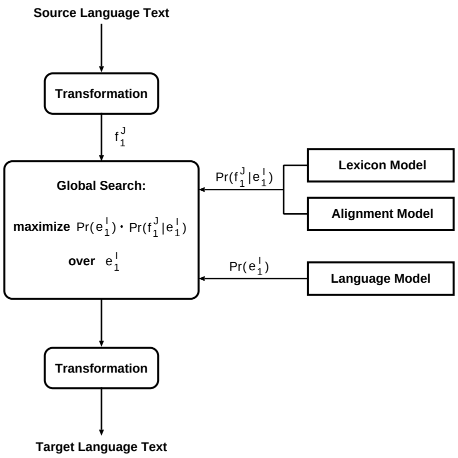

Figure 2: Architecture of the Translation Approach based on Bayes decision rule.

In our translation tool, we use a translation model that slightly differs from the model introduced as *Model 2* in [3]. It is based on a decomposition of the joint probability for f_1^J into a product of the probabilities for each word f_j:

$$Pr(f_1^J|e_1^I) = p(J|I) \cdot \prod_{j=1}^{J} p(f_j|e_1^I) \ , \tag{2}$$

where the lengths of the strings are regarded as random variables and modelled by the distribution $p(J|I)$. Now we assume a sort of pairwise interaction between the input word f_j and each output word e_i in e_1^I. These dependencies are captured in the form of a mixture distribution:

$$p(f_j|e_1^I) = \sum_{i=1}^{I} p(i|j,J,I) \cdot p(f_j|e_i) \ . \tag{3}$$

Inserting this into (2), we get

$$Pr(f_1^J|e_1^I) = p(J|I) \prod_{j=1}^{J} \sum_{i=1}^{I} p(i|j,J,I) \cdot p(f_j|e_i) \tag{4}$$

with the following components: the sentence length probability $p(J|I)$, the alignment probability $p(i|j,J,I)$ and the lexicon probability $p(f|e)$.

All these probabilities are estimated fully automatically from huge corpora of parallel texts.

The decoder used in the translation tool is based on Dynamic Programming with sequential expansion of the output sequence, as described in [8].

2 Approach to Sign Language Translation System

In the previous section, the basic theory for both modules, the recognition part and the translation part, has been introduced. This section details the approach of an automatic translation system composed of the two seperated modules. The general idea of the translation system is illustrated in figure 3.

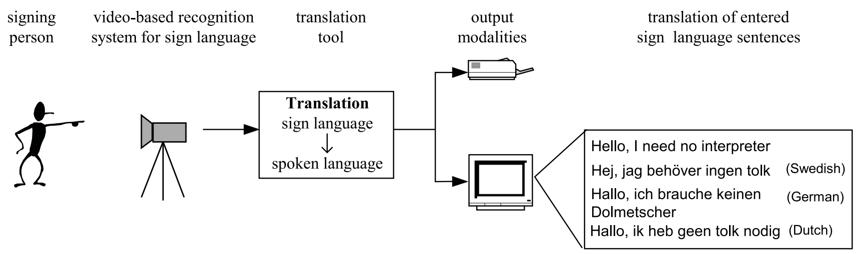

Figure 3: Components of the video-based continuous sign language translation system

Sign languages are visual languages and they can be characterised by manual (handshape, handorientation, location, motion) and non-manual (trunk, head, gaze, facial expression, mouth) parameters [2]. One-handed and two-handed signs are used. For one-handed signs only the so called dominant hand performs the sign. For two-handed signs both hands of the signer make the sign (symmetrical/non-symmetrical); the second hand is called non-dominant hand.

After recording, the sequence of input images is digitised and segmented. In the next processing step features regarding size, shape and position of the fingers, hands and body of the signer are calculated. Using this information a feature vector is built that reflects the manual sign parameters. Manual sign parameters are for example: relative position of the dominant and non-dominant hand, the size of of each finger, as well as palm and back of the dominant hand and the total size of the non-dominant hand.

These vectors serves as input for the sign language recognition system. Classification is performed by using HMMs. For both, training and recognition, feature vectors must be extracted from each video frame and input into the HMM. After the classification of the continuous signed sentence, the stream of recognised signs is inputted into the translation tool. Here, this stream is converted into a meaningful sentence of spoken language with a proper grammatical structure. It is not mandatory to have the same spoken as signed language. Since the translation tool is based on statistical methods a translation i.e. from German Sign Language into english is thinkable. Here, all sentences to be translated are not restricted and do not have to be defined in advance in a developed dictionary, such as the dialogue system for japanese sign language of Lu [6].

3 Evaluation of the System

Up to date the complete translation system is not yet tested. Results exist for both modules separately, the video-based continuous sign language recognition tool and the translation tool applied to three different translation tasks of variable complexity.

3.1 Experiments for the recognition tool

The vocabulary of the sign database consists of 100 signs representing different word types, such as nouns, verbs, adjectives, etc. The signs were chosen from the domain *Shopping at the Supermarket*. Since a critical issue in HMM-based language recognition systems is training data, we collected 6 hours of training and 1 hours of test data on a video tape. The system is trained and tested by one person. The native language of the signing person is german. The person is working as an interpreter for GSL and therefore did not learn the signs explicitely for this task.

Preparing the training set, we focused on the construction of sign sentences with different sign order. It is important to mention that the independent test set includes sign successions which are not part of the training set. No intentional pauses are placed between signs within a sentence, but the sentences themselves are separated. Constraints regarding a specific sentence structure are not allowed. The avoidance of minimal pairs is not an aim. A minimal pair is a pair of signs that differ in only one parameter [12]. All sentences of the sign database are meaningful and grammatically well-formed. Each sentence ranges from two to nine signs in length [1].

The experiments carried out are based on signer-dependent recognition. The sign recognition results are illustrated in table 1.

Table 1: Sign accuracy for differents vocabularies

Vocabulary	Accuracy
52 Signs	94.0 %
100 Signs	91.6 %

As can be seen from table 1, sign accuracy is ranging from 91.6% to 94.0% depending on the size of the vocabulary.

Analysing the results, it can be stated that the system is able to recognise continuous sign language. Considering a lexicon of 100 signs the system achieves an accuracy of 91.6%. Looking closer at the results it is obvious that the system discriminates most of the minimal pairs. Another important aspect is the fact that the unseen sign transitions in the test set are recognised in a good manner. Furthermore, the achieved recognition performance indicates that the system is able to handle the free order of signs within a sentence.

3.2 Experiments for the translation tool

By the time of submission, the translation tool has not yet been tested on the output of the sign recognition tool as input to the translation decoder. Results will be present by the time of the final submission of the camera ready version.

There have been various experiments with different complexity:

The Feldmann-Corpus is very simple task. The input sentences are descriptions of geometrical scenes in Spanish. The input language vocabulary consists of 37 words. The output sentences are generated in English using 28 different words. Automatic translation is performed with 100% accuracy.

The EuTrans-Corpus is a Spanish-to-English corpus of sentences from the touristic domain. Vocabulary sizes are 686 for the input language and 513 for the output language. Translation accuracy is 88%.

The Verbmobil-Corpus is a German-to-English corpus with spoken dialogues from the appointment scheduling domain. Here, the syntactic structures of the sentences are highly variable due to effects of spontaneous speech. The vocabulary is comparatively large (5936 words in German and 3505 words in English). Translation accuracy drops to 47%.

Comparing these results and given the vocabulary size of 100 different sign identifiers in the sign language translation task, we expect the automatic translation tool to perform very well (i.e. accuracy better than 90%) on this task.

4 Summary

In this paper we introduced two main modules to develop in the near future an automatic sign language translation system. The HMM-based continuous sign language recognition system is equipped with a single colour video camera for image recording. The extracted sequence of feature vector reflects the manual sign parameters. Visual modelling is carried out using HMMs, where each sign is modelled by a single HMM. Heading for user-dependent recognition, the recognition tool achieves a sign accuracy of 91.6%, based on a lexicon of 100 signs of German Sign Language. The sequence of recognised signs is passed to the translation tool, which converts this stream into a meaningful sentence of spoken language with a proper grammatical structure. The translation tool is expected to achieve a word accuracy of more than 90%. By the time of the final submission first results of the overall system will be presented.

References

[1] Bauer, B.: *Videobasierte Erkennung kontinuierlicher Gebärdensprache mit Hidden Markov Modellen.* Diploma Thesis, Aachen University of Technology (RWTH), Department of Technical Computer Science, 1998.

[2] Boyes Braem, P.: *Einführung in die Gebärdensprache und ihre Erforschung.* Signum Press, Hamburg, 1995.

[3] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer: The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311, 1993.

[4] Hienz, H., K.–F. Kraiss, and B. Bauer: *Continuous Sign Language recognition using Hidden Markov Models* In Tang, Y. (Ed.): ICMI'99 – The Second International Conference on Multimodal Interface, pp. IV10 – IV15, Hong Kong (China), 1999.

[5] Hienz, H., B. Bauer, and K.–F. Kraiss: *HMM-Based Continuous Sign Language Recognition using Stochastic Grammars* In: GW'99 – The 3rd Gesture Workshop: Towards a Gesture-Based Communication in Human-Computer Interaction, Gif-sur-Ivette (France), 1999.

[6] Lu, S., S. Igi, H. Matsuo, and Y. Nagashima: *Towards a Dialogue System Based on Recognition and Synthesis of Japanese Sign Language* In Wachsmuth, I. and M. Fröhlich (Editors): Gesture and Sign Language in Human Computer Interaction, International Gesture Workshop Bielefeld 1997, pp. 273–284, Bielefeld (Germany), Springer, 1998.

[7] Braffort, A.: *ARGo: An Architecture for Sign Language Recognition and Interpretation.* In P. Harling and A. Edwards (Editors): Progress in Gestural Interaction, pp. 17–30, Springer, 1996.

[8] S. Niessen, S. Vogel, H. Ney, C. Tillmann: A DP-Based Search Algorithm for Statistical Machine Translation. *Proc. 35th Annual Conference of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pp. 960–967, Montreal, Canada, August 1998.

[9] Rabiner, L.R. and B.H. Juang: *An Introduction to Hidden Markov Models.* In IEEE ASSP Magazin, pp. 4–16, 1989.

[10] Schukat-Talamazzini, E.G: *Automatische Spracherkennung.* Vieweg Verlag, 1995.

[11] Starner, T., J. Weaver and A. Pentland: *Real-Time American Sign Language Recognition using Desk- and Wearable Computer-Based Video.* In IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12):1371–1375, 1998.

[12] Stokoe, W., D. Armstrong and S. Wilcox: *Gesture and the Nature of Language.* Cambridge University Press, Cambridge (UK), 1995.

[13] Vogler, C. and D. Metaxas: *Adapting Hidden Markov Models for ASL Recognition by using Three-Dimensional Computer Vision Methods.* In Proceedings of IEEE International Conference and Systems, Man, and Cybernetics, pp. 156–161, Orlando (USA), 1997.