Continuous Sign Language Recognition – Approaches from Speech Recognition and Available Data Resources

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

Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University
D–52056 Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

In this paper we describe our current work on automatic continuous sign language recognition. We present an automatic sign language recognition system that is based on a large vocabulary speech recognition system and adopts many of the approaches that are conventionally applied in the recognition of spoken language. Furthermore, we present a set of freely available databases that can be used for training, testing and performance evaluation of sign language recognition systems. First results on one of the databases are given, we show that the approaches from spoken language recognition are suitable, and we give directions for further research.

1. Introduction

The first generation of sign language recognition systems has employed special data acquisition tools like gloves or wearable cameras to obtain the features to recognize the gestures (Vogler and Metaxas, 1997; Starner et al., 1998; Bauer et al., 2000). Only few research groups use databases which have been recorded using normal stationary cameras (Bowden et al., 2004; Zahedi et al., 2005; Zieren and Kraiss, 2005). Nonetheless, most of the databases have been recorded in a highly restricted environment with constant lightning, homogeneous, non-changing background, and the signers are dressed in long-sleeve shirts. In such an environment motion- and skin-color detection is greatly simplified, resulting in a task that is only slightly more difficult than the tasks where data-gloves were used.

Some other databases have been created by linguistic research groups. These databases have not been produced with sign language recognition in mind; i.e., no suitable transcription is available. To use these data for the training or for performance evaluation in sign language recognition systems, the necessary transcriptions have to be created, which is a costly process and requires a lot of human work.

In this paper, we present different databases which have been prepared in different ways: (i) I6-Boston201 database: consists of 201 sentences of American sign language (ASL) and have been recorded in a controlled environment. The signs have been recorded by four standard stationary cameras. It is a subset of the database recorded by Boston University (Neidle et al., 2000). (ii) Phoenix database: has been recorded from the daily news "Tagesschau" of the German TV channel Phoenix. In this program an interpreter signs the news in German sign language simultaneously in the lower right corner of the TV screen. This database is transcribed in German sign language and German language. The movies are not recorded in a controlled environment, but instead the signer is shown in front of a strongly non-homogeneous, non-constant background. (iii) the ECHO database consists of three corpora: British sign language (BSL) (Woll et al., 2004), Swedish sign language (SSL) (Bergman and Mesch, 2004) and sign language of the Netherlands (NGT) (Crasborn et al., 2004), respectively. We have prepared the ECHO databases for sign language recognition by choosing some parts of the original corpora and creating the necessary annotations.

Our automatic sign language recognition system is derived from a large vocabulary automatic speech recognition system, because both, speech and sign language are sequences of the features over the time. In section 2, a short overview of the system is presented. We will introduce the databases in section 3 and finally preliminary results of the system and conclusion are shown in section 4 and 5.

2. System Overview

As mentioned above, our sign language recognition system is based on a large vocabulary speech recognition system (Kanthak et al., 2000b; Gollan et al., 2005). This allows us to easily use the techniques developed for speech recognition and transfer the insights from this domain into automatic sign language recognition. Common speech recognition systems are based on the Bayes' decision rule,

$$\hat{w}_1^N = \arg\max_{w_1^N} \{Pr(w_1^N) \cdot Pr(x_1^T | w_1^N)\} \qquad (1)$$

where $\hat{w}_1^N$ is the sequence of words that is recognized, $Pr(w_1^N)$ is the language model, and $Pr(x_1^T|w_1^N)$ is the visual model (cp. acoustic model in speech recognition), x_1^T are the features for the time slots 1 to T.

Obviously, the features x_1^T have to be extracted in a different way than in speech recognition using techniques known from the image processing domain. To handle video files we use the FFmpeg library[1], which is able to handle a wide range of different video formats. Basic image processing methods are integrated into the system: thresholding, cropping, rotation, resizing to allow for a suitable selection of the region of interest in the videos; convolution, Sobel filters, smoothing to pre-process images. Furthermore, methods that were successfully used in gesture recognition were integrated: skin color models (Jones and Rehg, 1998) to locate faces and hands, motion detection

[1] http://ffmpeg.sourceforge.net/index.php

Figure 1: Sample frames from I6-Boston201.

by difference images (Dreuw et al., 2006), motion history images (Morrison and McKenna, 2004), geometric features (Rigoll et al., 1998), and spatial features (Bowden et al., 2004). In (Dreuw et al., 2006) a tracking algorithm using dynamic programming was introduced that considers the complete image sequence to find the best tracking-path with respect to a given criterion. This tracking can be used in the recognition process in the same way as time-alignment is used in speech recognition.
This framework allows for easy testing and development of new features for automatic sign language recognition. It is easily possible to reconfigure the system, to change parameters, to use different corpora and to change the feature extraction process. A description of the speech recognition system can be found in (Kanthak et al., 2000a).

3. Databases

In this section, three different sign language databases are presented. These databases are a starting point for performance evaluation in automatic sign language recognition. Where missing, we created the necessary annotation to be able to use them for automatic sign language recognition. All the data are freely available on the Internet.

3.1. I6-Boston201 Database

The National Center for Sign Language and Gesture Resources of the Boston University has published a database of ASL [2]. We have used 201 annotated videos of ASL sentences. Although this database was not recorded primarily for image processing and recognition research, we considered it as a starting point for a recognition corpus because the data are available to other research groups and can thus be a basis to compare different approaches. The database consists of videos from three signers: one male and two female signers. The signers are dressed differently.
The signing is captured simultaneously by four stationary standard cameras, three of them are black/white cameras and one is a color camera. All cameras have fixed positions. Two sample frames are shown in Figure 1.
Two black/white cameras, directed towards the signer's face, form a stereo pair that can be used to obtain three-dimensional data. Another camera is installed on the side of the signer.
The color camera is placed between the cameras of the stereo pair and is zoomed to capture only the face of the signer. This camera can be used for facial expression analysis. The movies are recorded at 30 frames per second and the size of the frames is 312×242 pixels. We use the published video streams at the same frame rate but extract the upper center part of size 195×165 pixels. (Parts of the bottom of the frames show some information about the frame, and the left and right border of the frames are unused.)
To use these data for ASL sentence recognition, we separated the recordings into a training and evaluation set. To optimize the parameters of the system, the training set is further split into separate training and development parts. To optimize parameters in the training process, the system is trained by using the training set and evaluated using the development set. When parameter tuning is finished, the training data and development data are used to train one model using the optimized parameters. This model is then evaluated on the so-far unseen test set. This database is called I6-Boston201 in the following. Corpus statistics for this database are shown in Table 1 which include number of sentences, running words, unique words and singletons in the each part. Singletons are the words occurring only once in the set.

	Training set		Evaluation set
	Training	Development	
Sentences	131	30	40
Running Words	695	172	216
Unique Words	103	65	79
Singletons	37	38	45

Table 1: Corpus statistics for I6-Boston201 database.

Figure 2: Left: whole screen image, right: close up to the interpreter.

3.2. Phoenix Database

The German TV channel Phoenix broadcasts the daily "Tagesschau" news program in German and with a German sign language translation in the lower right corner of the screen. The whole screen and a close up of the interpreter are shown in Figure 2. We have recorded the complete "Tagesschau" for 104 days and currently a snapshot of the recordings consisting of the weather reports of 51 days is used. The sign language of these recordings is fully transcribed. These data are split into training, development, and test data and the complete corpus statistics of this database is given in Table 2. In total there are 11 different signers (1 male and 10 females).
The movies are in MPEG1 video format and in PAL res-

[2] http://www.bu.edu/asllrp/ncslgr.html

	Set		
	Training	Development	Evaluation
Sentences	421	79	56
Running Words	5890	500	389
Unique Words	643	168	139
Singletons	0	70	63

Table 2: Corpus statistics for Phoenix database.

Figure 3: Sample frames from ECHO databases.

olution (352×288). The database transcription has been created by a congenitally deaf using the ELAN software[3]. In addition to the pure transcription, information on the signers, start time and end time of the gestures and also boundaries of the sentences are available in the annotation files. Further information about annotation is available in (Bungeroth et al., 2006).

3.3. ECHO-Databases

The ECHO database[4] consists of three corpora in BSL, SSL and NGT. All three corpora include the videos from sign narrations of the same five fable stories, a small lexicon and interviews with the signers. In addition, there is sign language poetry in BSL and NGT. Figure 3 shows sample image frames. The corpora have been annotated linguistically and include sign language and spoken language transcription in English. In addition, SSL and NGT sections include Swedish and Dutch transcription, respectively.
Also these videos have been transcribed using the ELAN software and the transcription includes word and sentence boundaries for the sign language recognition.
To use the ECHO databases in the field of sign language recognition, we have chosen some parts of the five fable stories of the original database and have created a database for each of the subcorpora. We name these databases ECHO-BSL, ECHO-SSL, ECHO-NGT.
Although the data have been recorded in a completely controlled environment with constant background, it is currently very hard to use these three databases for sign language recognition: The number of singletons and the number of unique words are too high in relation to the total number of utterances. To reduce the data sparseness, we have decided to split the corpus into training and testing data only, i.e. for these corpora no development sets have been specified. Furthermore, the test set was selected to have no out-of-vocabulary words, i.e. each word in the test set is at least once in the respective training set. The training corpora consists of the sentences and also segmented words of them but evaluation contains only sentences.

[3] http://www.mpi.nl/tools/elan.html
[4] http://www.let.ru.nl/sign-lang/echo

	Training set	Evaluation set
Sentences	206	56
Running Words	2628	237
Unique Words	534	97
Singletons	343	57

Table 3: Corpus statistics for ECHO-BSL database.

	Training set	Evaluation set
Sentences	136	23
Running Words	2988	129
Unique Words	520	70
Singletons	280	44

Table 4: Corpus statistics for ECHO-SSL database.

3.3.1. ECHO-BSL

The ECHO-BSL database is signed by 2 signers (1 male and 1 female). Statistics of the corpus is shown in Table 3.

3.3.2. ECHO-SSL

The ECHO-SSL database is signed by a male signer. Statistics of the corpus is shown in Table 4.

3.3.3. ECHO-NGT

The ECHO-NGT database is signed by 3 signers (2 males and 1 female). Statistics of the corpus is shown in Table 5.

4. Preliminary Results

In this section we present some preliminary results on the I6-Boston201 corpus introduced in the previous section. For the experiments, the video frames were scaled down to the size of 32×32 pixels. The performance of the system is measured by the word error rate (WER) which is equal to the number of deletion, substitution and insertion of the words divided by the number of running words. The results on development and evaluation sets including the perplexity (PP) and WER of the system using different language models are shown in Table 6. The n-gram language models where the probability of a sentence is estimated from the conditional probabilities of each word given the $n-1$ preceding words are employed in the experiments. The n-gram language models are called zerogram, unigram, bigram and trigram where n is equal to 0, 1, 2 or 3, respectively.

	Training set	Evaluation set
Sentences	187	53
Running Words	2450	197
Unique Words	468	77
Singletons	268	40

Table 5: Corpus statistics for ECHO-NGT database.

Language Model	Development set		Evaluation set	
	PP	WER(%)	PP	WER(%)
Zerogram	105	75	105	65
Unigram	36	71	37	63
Bigram	8	68	9	57
Trigram	7	69	6	55

Table 6: Preliminary result of the system on I6-Boston201.

Currently, we are working with the other corpora and we are trying to find a suitable set of image features for good recognition results. Furthermore the parameters of the sign language recognition system have to be tuned towards the task at hand as the parameters that are used in speech recognition are not always suited for the recognition of sign language.

5. Conclusion

We have presented an overview of our current efforts in the recognition of sign language. In particular we have employed a large-vocabulary speech recognition system which was extended by basic image processing techniques and which is currently being extended with feature extraction methods for sign language recognition. Furthermore, we presented 5 different tasks which can be used to benchmark continuous sign language recognition systems. These databases are freely available and can thus be used by other research groups.

Acknowledgments

The authors would like to thank Georg Heigold and András Zolnay for fruitful discussion and lots of tips on how to use the speech recognition system. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) under grand NE-572/6.

6. References

B. Bauer, H. Hienz, and K.F. Kraiss. 2000. Video-based continuous sign language recognition using statistical methods. In *Proceedings of the International Conference on Pattern Recognition*.

B. Bergman and J. Mesch, 2004. *ECHO Data Set for Swedish Sign Language (SSL)*. Department of Linguistics, University of Stockholm, http://www.let.ru.nl/sign-lang/echo.

R. Bowden, D. Windridge, T. Kabir, A. Zisserman, and M. Bardy. 2004. A linguaistic feature vector for the visual interpretation of sign language. In *Proceedings of ECCV 2004, the 8th European Conference on Computer Vision*.

J. Bungeroth, D. Stein, P. Dreuw, M. Zahedi, and H. Ney. 2006. A german sign language groups of the domain weather report. In *Proceedings of the 5th Iternational Conference on Language Resources and Evaluation*.

O. Crasborn, E. van der Kooij, A. Nonhebel, and W. Emmerik, 2004. *ECHO Data Set for Sign Language of the Netherlands (NGT)*. Department of Linguistics, Radboud University Nijmegen, http://www.let.ru.nl/sign-lang/echo.

P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. 2006. Tracking using dynamic programming for appearance-based sign language recognition. In *Proceedings of the 7th International Conference of Automatic Face and Gesture Recognition*.

C. Gollan, M. Bisani, S. Kanthak, R. Schlter, and H. Ney. 2005. Cross domain automatic transcription on the tc-star epps corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 825–828, Philadelphia, PA, March.

M. Jones and J. Rehg. 1998. Statistical color models with application to skin color detection. CRL 98/11, Compaq Cambridge Research Lab.

S. Kanthak, S. Molau, A. Sixtus, R. Schlüter, and H.Ney. 2000a. The rwth large vocabulary speech recognition system for spontaneous speech. In *Proceedings of the Konvens 2000*.

S. Kanthak, A. Sixtus, S. Molau, R. Schlter, and H. Ney, 2000b. *Fast Search for Large Vocabulary Speech Recognition*.

K. Morrison and S.J. McKenna. 2004. An experimental comparison of trajectory-based and history-based representation for gesture recognition. In *Proceedings of the International Gesture Workshop*.

C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. Cambridge, MA: MIT Press.

G. Rigoll, A. Kosmala, and S. Eickeler. 1998. High performance real-time gesture recognition using hidden markov models. In *Proceedings of Iternational Gesture Workshop*.

T. Starner, J. Weaver, and A. Pentland. 1998. Real-time american sign language recognition using desk and wearable computer based video. *Transaction of Pattern Analysis and Machine Intelligence*, 20(2):1371–1375.

C. Vogler and D. Metaxas. 1997. Adapting hidden markov models for asl recognition by using three-dimentional computer vision methods. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*.

B. Woll, Sutton-Spence, and D. Waters, 2004. *ECHO Data Set for British Sign Language (BSL)*. Department of Language and Communication Science, City University (LONDON), http://www.let.ru.nl/sign-lang/echo.

M. Zahedi, D. Keysers, T. Deselaers, and H. Ney. 2005. Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In *Proceedings of DAGM 2005, 27th Annual meeting of the German Association for Pattern Recognition*.

J. Zieren and K.F. Kraiss. 2005. Robust person-independent visual sign language recognition. In *Proceedings of the 2nd Iberian Conference on Pattern Recognition and Image Analysis*.