Benchmark Databases for Video-Based Automatic Sign Language Recognition.

Conference Paper · January 2008

Source: DBLP

CITATIONS
67

READS
516

5 authors, including:

Philippe Dreuw
Robert Bosch GmbH, Hildesheim, Germany
43 PUBLICATIONS **1,097** CITATIONS

SEE PROFILE

Stan Sclaroff
Boston University
338 PUBLICATIONS **17,405** CITATIONS

SEE PROFILE

Hermann Ney
RWTH Aachen University
987 PUBLICATIONS **38,005** CITATIONS

SEE PROFILE

Benchmark Databases for Video-Based Automatic Sign Language Recognition

Philippe Dreuw[1], Carol Neidle[2], Vassilis Athitsos[3], Stan Sclaroff[2], and Hermann Ney[1]

[1]RWTH Aachen University, Aachen, Germany
dreuw@cs.rwth-aachen.de
[2]Boston University, Boston, MA, USA
carol@bu.edu
[3]University of Texas, Arlington, TX, USA

Abstract

A new, linguistically annotated, video database for automatic sign language recognition is presented. The new RWTH-BOSTON-400 corpus, which consists of 843 sentences, several speakers and separate subsets for training, development, and testing is described in detail. For evaluation and benchmarking of automatic sign language recognition, large corpora are needed. Recent research has focused mainly on isolated sign language recognition methods using video sequences that have been recorded under lab conditions using special hardware like data gloves. Such databases have often consisted generally of only one speaker and thus have been speaker-dependent, and have had only small vocabularies. A new database access interface, which was designed and created to provide fast access to the database statistics and content, makes it possible to easily browse and retrieve particular subsets of the video database. Preliminary baseline results on the new corpora are presented. In contradistinction to other research in this area, all databases presented in this paper will be publicly available.

1. Introduction

Currently available sign language video databases were created for linguistic purposes or gesture recognition using small vocabularies. Most databases used in sign language processing so far do not provide or include what is important for the evaluation of sign language processing algorithms (Bowden et al., 2004; Martinez et al., 2002). An overview of available language resources for sign language processing and especially recognition is presented in (Zahedi et al., 2006). Recently an Irish sign language database has been released (Stein et al., 2007). Here we focus on benchmark databases that can be used for investigating linguistic problems, and evaluating automatic sign language recognition systems or statistical machine translation systems.

The National Center for Sign Language and Gesture Resources at Boston University has published an expanding database of American Sign Language (ASL). Dreuw and colleagues from the RWTH Aachen University created several subsets for the evaluation of isolated and continuous sign language recognition: RWTH-BOSTON-50, RWTH-BOSTON-104, and the new RWTH-BOSTON-400 (see Section 4.). The new RWTH-BOSTON-400 is the largest publicly available benchmark corpus for video-based continuous sign language recognition. It contains 843 sentences, several speakers, and separate splits for training, development, and testing of automatic sign language recognition systems.

The RWTH-BOSTON-400 database is created from a subset of the larger data set available through Boston University. The BU ASL corpus has been used previously in evaluation of computer vision and pattern recognition methods, including detection of head gestures (Erdem and Sclaroff, 2002), recognition of facial expressions (Vogler and Goldenstein, 2007), hand tracking and recognition of hand shapes and movements (Vogler and Metaxas, 2004; Yuan et al., 2005; Dreuw et al., 2006), as well as categorization of signs (Zahedi et al., 2005; Tsechpenakis et al., 2006). Both data sets can be used for the training and evaluation of sign language recognition algorithms, as well as development and testing of human-machine interface approaches, data mining, etc. Since the data sets include annotations of non-manual gestures and movements that are integral to sign language, these data sets could also be used as a resource for investigating correlations, temporal relations, and alignments across manual and non-manual gestural channels.

2. Multimodal Resources for ASL

The National Center for Sign Language and Gesture Resources (NCSLGR) at Boston University[1] has been engaged in the collection of ASL data (including sets of individual utterances, narratives, and dialogues) from Deaf native signers. The NCSLGR makes available high-quality video files showing the signing from multiple angles, including a close-up of the face, in a variety of video formats, along with linguistic annotations that have been carried out in conjunction with the American Sign Language Linguistic Research Project (ASLLRP) at Boston University, using SignStream™ (Neidle et al., 2001; Neidle, 2002a)[2]. The annotations, available as SignStream™ files and in XML format, include indication of the start and end points of linguistically significant behaviors, including individual signs, produced by the hands and arms, and facial gestures (e.g., eyebrow position, eye aperture) and head movements (including nods and shakes) that have grammatical significance[3]. The annotation conventions are documented in

[1]http://www.bu.edu/asllrp/cslgr/

[2]A Macintosh Classic application designed by this group to facilitate linguistic analysis of visual language data. A Java reimplementation currently underway will contain many new features, including tools for fine-grained phonological annotation.

[3]The research described here that has been carried out at Boston University has been supported in part by grants from the National Science Foundation (HCC- 0705749, CNS-04279883,

(Neidle, 2002b; Neidle, 2007) and available for download[4]. These data are distributed in a couple of ways: (1) The video files and SignStream™ annotations are available on CD-ROM; six new CD-ROMs, containing 15 short narratives plus over 200 additional utterances, were released in August of 2007[5]. The 7 CD-ROMs currently being distributed include a total of over 1300 linguistically annotated utterances. (2) There is also a new Web interface to facilitate search of the existing data and download of subsets of video files and corresponding annotations that may be of interest to researchers, which is described in the next section.

3. Database Access Interface

The Database Access Interface (DAI)[6] is implemented in PHP/MySQL and runs in any modern browser. The interface allows users to query the data (or some user-specified subset of the data) in search of specific signs (or types of signs, e.g. fingerspelled signs), non-manual behaviors, or combinations thereof, while facilitating transfer of video files and annotations from the web site to the user's computer without the need of thirdparty software. The annotations are available either as SignStream™ files or as XML.

Video files are available in a variety of formats that offer different trade-offs between file size and video quality. The original, uncompressed video sequences have resolution of 648x484 pixels, and were recorded at 60 frames per second. Grayscale and 1CCD color cameras were used for recording the sequences. Bayer interpolation must be performed to convert the raw 1CCD output to RGB color. Each sequence was captured simultaneously by multiple (two to four) synchronized cameras: one or two cameras showing a front view of the upper body of the signer, one camera zooming in on the face from the front, and in many cases a camera showing the signer's upper body from the side. Calibration sequences are available for most of the recording sessions. The calibration sequences show a chessboard-like calibration pattern at a variety of 3D orientations, as seen from multiple cameras.

Compressed video files are also available for each sequence. The compressed files are in QuickTime format, and use Sorenson encoding. The image resolution in the compressed sequences is 324x242, and the frame rate is 30 frames per second. Each frame of a compressed sequence contains, in addition to the image data, a black-and-white field of size 324x70 at the bottom (for a total frame size of 324x312). The black-and-white field displays the date and time that the frame was captured, the frame number, the number of milliseconds from the beginning of the recording, and a frame ID that is unique across all recorded videos. This unique frame ID is also encoded in binary representation using black and white squares at the bottom of the image. Frame IDs are assigned sequentially. The information in this field can be used to verify frame correspondences between compressed and uncompressed sequences. The uncompressed sequences include this information in a black-and-white field at the bottom, or in a text stream that is part of the AVI file, or in an accompanying small binary file (with extension .time instead of .avi), that is available for downloading from the same directory as the uncompressed sequences.

Listing 1: Shortened XML annotation structure

```
<?xml version="1.0" encoding="UTF-8"?>
<CODING-SCHEME>
 ...
<FIELD ID="20001" NAME="English translation">
<FIELD ID="10000" NAME="main gloss">
 ...
</CODING-SCHEME>
 ...
<UTTERANCES>
  <UTTERANCE ID="0" S="166" E="2700">
    ...
    <SEGMENT PARTICIPANT-ID="0" PRIMARY="false">
    <TRACK FID="10000">
      <A S="434" E="767">fs-JOHN</A>
      <A S="934" E="1034">FINISH</A>
      <A S="1367" E="1667">READ+</A>
      <A S="1900" E="2167">BOOK+</A>
    </TRACK>
     ...
    <TRACK FID="20001">
      <A S="0" E="2567">John finished reading the book.</A>
    </TRACK>
   </SEGMENT>
  </UTTERANCE>
   ...
</UTTERANCES>
```

The DAI tool web site also provides statistics about the annotated video data set, including the frequency of signs, non-manual gestures, etc. For example, it is possible to search for signs with at least a certain number of tokens in the data set, and then to retrieve all utterances (video and annotations) in which such a sign occurs. The same is true for searches over non-manual gestures. This search functionality can be used in retrieving a subset of the data that has a sufficient number of examples per class, which can be essential to support training and testing of machine learning methods for ASL recognition.

4. Databases

All databases presented in this section are freely available for further research in linguistics[7] and recognition[8]. The data were recorded by Boston University, the database subsets were defined at the RWTH Aachen University in order to build up benchmark databases that can be used for the automatic recognition of isolated and continuous sign language, respectively.

In the following we briefly describe some commonly used statistical measures w.r.t. automatic recognition:

running words are the total number of words in the corpus

unique words determine the vocabulary size

singletons are words (or word tuples) that occur only once

zerogram-, unigram-, bigram-, trigram- language models describe different linguistic contexts

4.1. RWTH-BOSTON-50

The RWTH-BOSTON-50 database was created for the task of isolated sign language recognition (Zahedi et al., 2005; Zahedi et al., 2006). It has been used for nearest-neighbor leaving-one-out evaluation of isolated sign language words.

IIS-0329009, EIA-9809340, IIS-9528985).

[4] http://www.bu.edu/asllrp/reports.html

[5] http://www.bu.edu/asllrp/cd/

[6] http://ling.bu.edu/asllrpdata/queryPages/

[7] http://www.bu.edu/asllrp/

[8] http://www-i6.informatik.rwth-aachen.de/aslr/

Table 1: RWTH-BOSTON-104 corpus statistics

	Training	Evaluation
sentences	161	40
running words	710	178
vocabulary	103	65
singletons	27	9
OOV	-	1
images	12422	3324

Table 2: RWTH-BOSTON-400 person statistics for training set

speaker name	segments	time [sec]
Ben	90	283.3s
Norma	142	375.267s
Mike	364	1219.77s
Lana	37	162.367s

It consists of 50 isolated words (83 distinct words when pronunciation information is included), 483 word utterances, and 2 speakers (1 female, 1 male).

4.2. RWTH-BOSTON-104

Recently, this database has been extended by Dreuw and colleagues with pronunciation information (see Section 5.2.), and has been used successfully for continuous sign language recognition experiments (Dreuw et al., 2007a). The corpus statistics and language model perplexities are shown in Table 1 and Table 4.

4.3. RWTH-BOSTON-400

Many different information tiers that are of interest for linguistic research are available in the database XML annotations. From the recognition point of view, we have focused so far on only two field ids in order to build a subset which can be used for automatic recognition and statistical machine translation, namely the *main gloss* field (id=10000) and the English translation field (id=20001) (see Listing 1).

The complete corpus currently contains 843 sentences in total and is divided into 633 train sentences, 106 development sentences, 104 evaluation sentences. There are 7768 running words, where the main vocabulary has a size of 406 words without pronunciation information (i.e. the main evaluation tokens) (see Table 3).

There are four signers who occur in the video database: 2 male speakers account for 454 segments, and 2 female speakers account for 179 segments (see Table 2).

Table 3: RWTH-BOSTON-400 corpus statistics

	Training	Development	Evaluation
sentences	633	106	104
running words	5733	678	589
vocabulary	483	74	36
singletons	217	10	2
OOV	-	7	0
images	49486	10016	9053

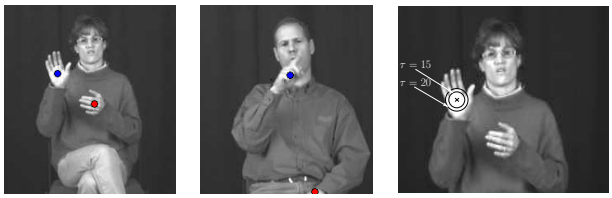

Figure 1: Sample frames of the RWTH-BOSTON-Hands database with annotated hand positions. Left and right hand are marked with red and blue circles respectively. The last image shows different tolerance radii $\tau = 15$ and $\tau = 20$.

4.4. RWTH-BOSTON-Hands

For the evaluation of hand tracking methods in sign language recognition systems a database has been prepared. The RWTH-BOSTON-Hands database is of a subset of the RWTH-BOSTON-104 videos with additional annotation of the signer's hand positions. The positions of both hands have been annotated manually in 15 videos. 1119 frames in total are annotated.

5. System Overview & Features

We give a short overview of the recognition framework and the used features presented in (Dreuw et al., 2007a).

5.1. Visual Modeling

Phonological analysis going back to (Stokoe et al., 1965) 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).

However, it is still unclear how best to approach recognition of these articulatory parameters. Although phonemes in spoken language are sequential, notwithstanding co-articulation 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 the main stream only. For this reason, we have relied on *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.

Appearance-Based Features. In our baseline system we use only appearance-based image features, 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.

They give a global description of all (manual and non-manual) features that have been shown to be linguistically important. The baseline system is Viterbi trained and uses a trigram LM (c.f. Section 5.3.).

Manual Features. To extract manual features, the dominant hand (i.e., the hand that is mostly used for one-handed signs and as finger spelling) is tracked in each image sequence. Given the hand position $u_t = (x, y)$ at time t in signing space, features such as hand velocity $m_t = u_t - u_{t-\delta}$ can easily be extracted. The hand trajectory features (Dreuw et al., 2007a) are similar to the features presented in (Vogler and Metaxas, 2001).

5.2. Pronunciation Handling

Signed languages exhibit dialectal variation comparable to that found in spoken languages. Thus, one may find signs quite different from one another (in all aspects of their articulation, including potentially duration) used by different signers with comparable meanings. For example, there are 5 different dialectal variants for "bread" used in Switzerland, and as many different ASL signs for "birthday" used in America.

It is also possible that a given sign may be produced with variations in articulation. For example,`GIVE`, `(2h)GIVE`, `(2h)alt.GIVE`: Each of these signs differ in shape or articulation, and should have a unique gloss model. Due to the relatively large vocabulary with many singleton observations, we put words with variations in articulation but with same meaning into the same evaluation class (i.e., they will have the same evaluation token). This helps us to estimate better language models and reduces the perplexity (see Section 5.3.).

Other problematic sources of variation include the natural variation that occurs for an individual speaker in multiple productions of the same sign, resulting sometimes from differences in linguistic context (giving rise to different co-articulation effects), speed of speech, slight variants in handshape allowed for particular signs, etc.

Small differences between the appearance and the length of the utterances are compensated for by the HMMs, but different pronunciations of a sign must be modeled by separate models, i.e. a different number of states and GMMs.

5.3. Language Models

A trigram LM was trained on the main gloss annotations of the training corpora using the SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney discounting with interpolation.

As explained in Section 4.3. for the RWTH-BOSTON-400 database, we train 483 word classes but we have only 400 evaluation tokens due to different pronunciations. Therefore, we have only a perplexity (PP) of 400 for a zerogram LM in Table 5, as the pronunciations are not considered in LM training (i.e. only the evaluation tokens, see Section 5.2.). A small perplexity corresponds to strong language model restrictions.

Table 4: RWTH-BOSTON-104 language model perplexities

LM type	Test PP
zerogram	106.0
unigram	36.8
bigram	6.7
trigram	4.7

Table 5: RWTH-BOSTON-104 language model perplexities

LM type	Development PP	Test PP
zerogram	400	400
unigram	63.4	50.9
bigram	32.3	26.2
trigram	30.1	25.1

6. Experimental Results

For the recognition of isolated signs we used the RWTH-BOSTON-50 database, where the currently best known word-error-rate of 17.2% WER has been reported in (Zahedi et al., 2005). Figure 2 shows the effect of using different n-gram language models and scales on the RWTH-BOSTON-104 database (Dreuw et al., 2007a). As in ASR, the language model adaptation by using sign language pronunciations achieves large improvements (the currently best known word-error-rate is 17.8% WER). Interestingly, the improvement factors achieved are similar to those from speech recognition (Klakow and Peters, 2002). Preliminary results for statistical machine translation on the recognizer output have been presented in (Stein et al., 2007). We achieve a 2.30% tracking error rate for a 20×20 search window on the RWTH-BOSTON-Hands database (Dreuw et al., 2006).

Movement epenthesis refers to movements that occur in natural sign languages when the location in the signing space changes between one sign and the next. In this work, Dreuw et al. added special labels in the RWTH-BOSTON-400 database for some of the movements that can occur in between signs (e.g. `[UP]`, `[DOWN]`, `[SILENCE]`) and also noted `[HOLD]`). Special care must be taken for LM handling of such movements, as they can be important indicators; for example, there is often a `[HOLD]` with the

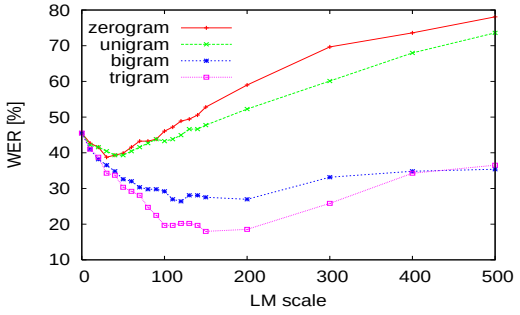

Figure 2: Results for different LMs and scales on the RWTH-BOSTON-104 database using the SRILM toolkit.

final sign in a question. Furthermore, sentence boundaries and unknown words are also labeled and handled by (Kneser and Ney, 1995) discounting with interpolation in the LM. Preliminary results for movement epenthesis in another subset of the database have been presented in (Yang et al., 2007).

With pronunciation information, movement epenthesis information, and sentence boundary information, the vocabulary to be trained consists of 482 words.

In training, 475 words from the complete vocabulary are trained, which results in an out-of-vocabulary (OOV) rate of 7 words. These words cannot be recognized using a whole-word model approach. After pronunciation handling in the training and recognition lexicon, we have in total 23 OOVs in the development corpus and 0 OOV words in the evaluation corpus.

For training the models, it is important to have a one-to-one correspondence between sign production and gloss. Encoding of pronunciation information makes it possible to distinguish variants. For the earlier RWTH-BOSTON-104 corpus annotations, Dreuw and colleagues added pronunciation information to the (preliminary version of the) Boston annotations and adapted the language models accordingly (Dreuw et al., 2007a). The new RWTH-BOSTON-400 corpus makes use of the newly released, fully verified, Boston NCSLGR annotations (Neidle, 2002b; Neidle, 2007). (The final verification of those annotations took some time, largely because of the importance of enforcing such one-to-one sign-gloss correspondences.)

Preliminary results on the large RWTH-BOSTON-400 database showed a couple of problems. One of the first steps in setting up an ASLR system from scratch is the start-stop silence detection and linear segmentation of the sentences in order to initialize the models to be trained. It turned out that opposed to ASR, where usually the energy of the audio signal is used for silence detection in the sentences (i.e. the corresponding motion in a video signal), 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 reliably speed up and automate the training process in order to improve the recognition performance.

Furthermore, there is also the issue of canonically one-handed vs. two-handed signs sometimes being produced with a non-canonical number of hands. Due to the lack of data, not all pronunciation models can be robustly estimated. In terms of recognition performance or word error rates, the confusion between such a one-handed version and a two-handed version would cause a substitution error, which can be suppressed by the use of evaluation tokens (c.f. Section 5.2.): the confusion of signs of the same class (i.e. meaning) are not considered as an error.

For signs that are normally produced as one-handed but that may also have two-handed realizations (or vice versa: canonically two-handed signs that may sometimes be produced with only 1 hand), the one-handed and two-handed versions can be trained independently, but use a common evaluation token.

Figure 3: Example of the four speakers: due to the different clothing (short sleeves, long sleeves, glasses, etc.) and camera setups, nine speaker setups have to be handled in the RWTH-BOSTON-400 database.

Another problem arises from the increased number of speakers and different environment setups: the usage of appearance-based full-body features for a baseline system as proposed in (Dreuw et al., 2007a) is of course limited to either many training examples or similar appearance of the speaker. Opposed to the simpler speaker setup of the RWTH-BOSTON-104 with only 3 speaker setups, we have to deal with at least 9 speaker setups in the RWTH-BOSTON-400 database (see Figure 3).

7. Summary & Conclusion

The largest publicly available benchmark corpus for continuous sign language recognition was presented. Promising results on the publicly available benchmark database RWTH-BOSTON-104 have been achieved for automatic recognition (Dreuw et al., 2007a) and translation (Dreuw et al., 2007b; Stein et al., 2007) that can be used as baseline reference for other researchers. However, the preliminary results on the larger RWTH-BOSTON-400 database show the limitations of the proposed framework and the need for better visual features and models.

In the future, the new database access will be used to further enrich the RWTH-BOSTON-400 annotations, and will open up the path to multiple stream processing (i.e. the independent recognition of hands, faces, body, ...) with a late fusion of the independent systems.

With the new database access interface it will be easier in the future to provide further benchmark subsets focussing on different problems in automatic recognition (such as speaker independence or part of speech tagging).

In the future, we can investigate which image features or which properties of the language model cause mistakes,

and which additional features, pronunciations, speaker or language model adaptations can remedy certain problems.

There is also an ongoing effort at Boston University to create a video sign lexicon with over 3,000 signs (including most of those contained in (Valli, 2006)), as produced by multiple signers. These video examples will be added to the corpus publicly accessible through the DAI. Our goal for the longer term is to enhance the DAI interface to support searches based on linguistic properties of signs, such as hand shape, once linguistic annotations including such information become available. Another potential enhancement to the DAI would be to support "query by example", where a user-provided video clip of a sign can be used to retrieve similar items (or utterances that contain similar items). Thus, the research in computer vision sign language recognition could be used to enhance the DAI search capabilities.

8. References

R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. 2004. A linguistic feature vector for the visual interpretation of sign language. In *European Conf. Computer Vision*, volume 1, pages 390–401.

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

P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. 2007a. Speech recognition techniques for a sign language recognition system. In *Interspeech 2007*, pages 2513–2516, Antwerp, Belgium, August. ISCA best student paper award of Interspeech 2007.

P. Dreuw, D. Stein, and H. Ney. 2007b. Enhancing a sign language translation system with vision-based features. In *International Workshop on Gesture in Human-Computer Interaction and Simulation*, pages 18–20, Lisbon, Portugal, May.

U. M. Erdem and S. Sclaroff. 2002. Automatic detection of relevant head gestures in American Sign Language communication. In *International Conf. on Pattern Recognition (ICPR)*, volume 1, pages 460–463.

D. Klakow and J. Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38:19–28.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *IEEE ICASSP*, volume 1, pages 49–52, Detroit, MI.

A. M. Martinez, R. B. Wilbur, R. Shay, and A. C. Kak. 2002. Purdue RVL-SLLL ASL database for Automatic Recognition of American Sign Language. In *IEEE Int. Conf. on Multimodal Interfaces*, Pittsburg, PA, USA, October.

C. Neidle, S Sclaroff, and V. Athitsos. 2001. Signstream^{TM}: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, and Computers*, 3(33):311–320.

C. Neidle. 2002a. Signstream^{TM}: A database tool for research on visual-gestural language. *Journal of Sign Language and Linguistics*, 1/2(4):203–214.

C. Neidle. 2002b. Signstream^{TM}annotation: Conventions used for the American Sign Language Linguistic Research Project. Technical Report 11, Boston University.

C. Neidle. 2007. Signstream^{TM}annotation: Addendum to conventions used for the American Sign Language Linguistic Research Project. Technical Report 13, Boston University, August.

D. Stein, P. Dreuw, H. Ney, S. Morrissey, and A. Way. 2007. Hand in hand: Automatic sign language to speech translation. In *11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 214–220, Skövde, Sweden, September.

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

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *ICSLP*, volume 2, pages 901–904, Denver, CO, September.

G. Tsechpenakis, D. Metaxas, and C. Neidle. 2006. Learning-based dynamic coupling of discrete and continuous trackers. *Computer Vision and Image Understanding*, 104(2-3):140–156, December.

C. Valli, editor. 2006. *The Gallaudet Dictionary of American Sign Language*. Gallaudet U. Press, Washington, DC, USA.

C. Vogler and S. Goldenstein. 2007. Facial movement analysis in ASL. *Springer Journal on Universal Access in the Information Society*, page to appear.

C. Vogler and D. Metaxas. 2001. A framework for recognizing the simultaneous aspects of American Sign Language. *Computer Vision & Image Understanding*, 81(3):358–384, March.

C. Vogler and D. Metaxas. 2004. Handshapes and movements: Multiple-channel ASL recognition. *Springer Lecture Notes in Artificial Intelligence*, (2915):247–258.

R. Yang, S. Sarkar, and B. Loeding. 2007. Enhanced level building algorithm for the movement epenthesis problem in sign language recognition. In *Computer Vision and Pattern Recognition*.

Q. Yuan, S. Sclaroff, and V. Athitsos. 2005. Automatic 2d hand tracking in video sequences. In *IEEE Workshop on Applications of Computer Vision*.

M. Zahedi, D. Keysers, and H. Ney. 2005. Pronunciation clustering and modeling of variability for appearance-based sign language recognition. In *International Gesture Workshop 2005*, volume 3881, Vannes, France, May.

M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, and H. Ney. 2006. Continuous sign language recognition - approaches from speech recognition and available data resources. In *Second Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, pages 21–24, Genoa, Italy, May.