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
[2]Boston University, Boston, MA, USA
[3]University of Texas, Arlington, TX, USA

Introduction

- currently available sign language video databases
  - for linguistic purposes
  - gesture recognition using small vocabularies
- here: new benchmark databases for evaluation of
  - linguistic problems
  - automatic sign language recognition
  - statistical machine translation

Multimodal Resources for ASL

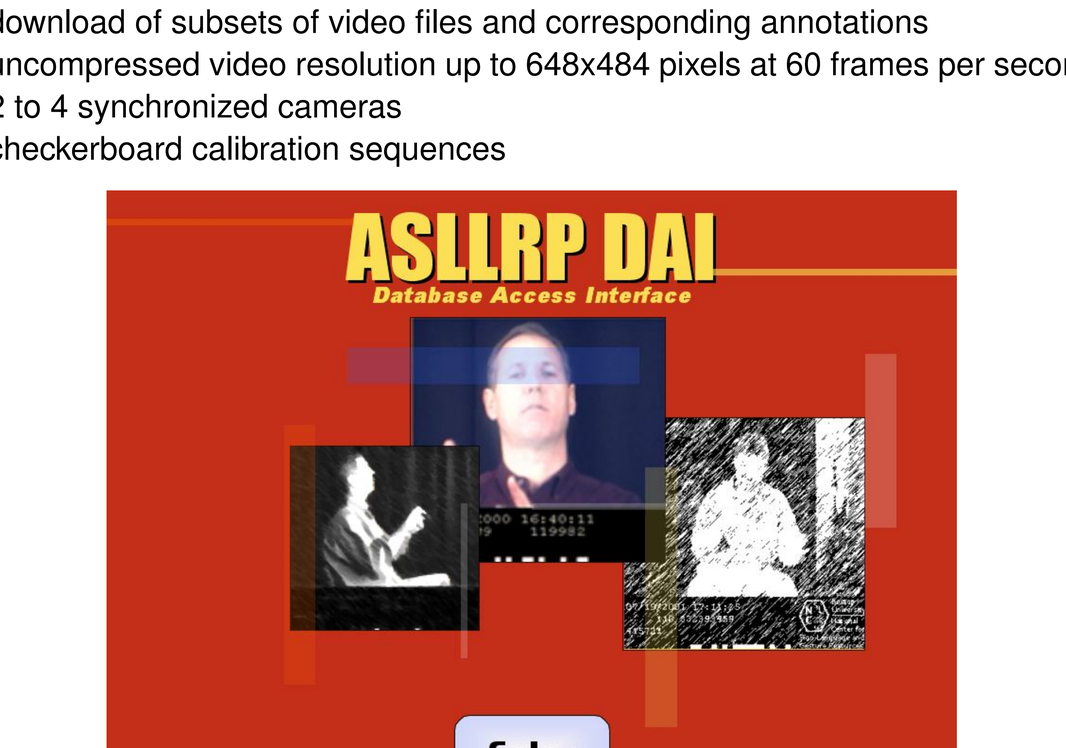
- National Center for Sign Language and Gesture Resources (NCSLGR) at Boston University
  - `http://www.bu.edu/asllrp/cslgr/`
- collection of American Sign Language data from deaf native signers
- high-quality video files in a variety of video formats
  - multiple angles
  - close-up of the face
  - with linguistic annotations

Linguistic Annotations

- American Sign Language Linguistic Research Project (ASLLRP)
- SignStream™ annotation software: `http://www.bu.edu/asllrp/`
- annotation format includes
  - indication of the start and end points of linguistically significant behaviors
  - individual signs, produced by the hands and arms
  - facial gestures (e.g. eyebrow position, eye aperture)
  - head movements (including nods and shakes) that have grammatical significance
- 7 CD-ROMs include a total of over 1300 linguistically annotated utterances
- available in SignStream™ or simple XML format

Database Access Interface

- search of the existing data
- download of subsets of video files and corresponding annotations
- uncompressed video resolution up to 648x484 pixels at 60 frames per second
- 2 to 4 synchronized cameras
- checkerboard calibration sequences



RWTH-BOSTON-50 Database

- 483 utterances of isolated words
- vocabulary size of 50 words, 83 with pronunciations
- 3 signers

RWTH-BOSTON-104 Database

- 201 utterances of continuous sign language sentences
- 3 signers
- 26% of the training data are singletons

- corpus statistics

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

- language model perplexities

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

- best known result is 12.9% WER

RWTH-BOSTON-400 Database

- corpus statistics

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

- language model perplexities

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

- person statistics for training set

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

- difficulties in preliminary results:
  - silence handling
  - movement epenthesis
  - canonically one-handed vs. two-handed signs
  - pronunciations
  - increased number of speakers

- several speaker setups

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

RWTH-BOSTON-Hands Database

- database with annotated hand and head positions

WWW

- freely available for further research in
  - linguistics:
    `http://www.bu.edu/asllrp/`
  - computer science:
    `http://www-i6.informatik.rwth-aachen.de/aslr/`