Enhanced Continuous Sign Language Recognition using PCA and Neural Network Features

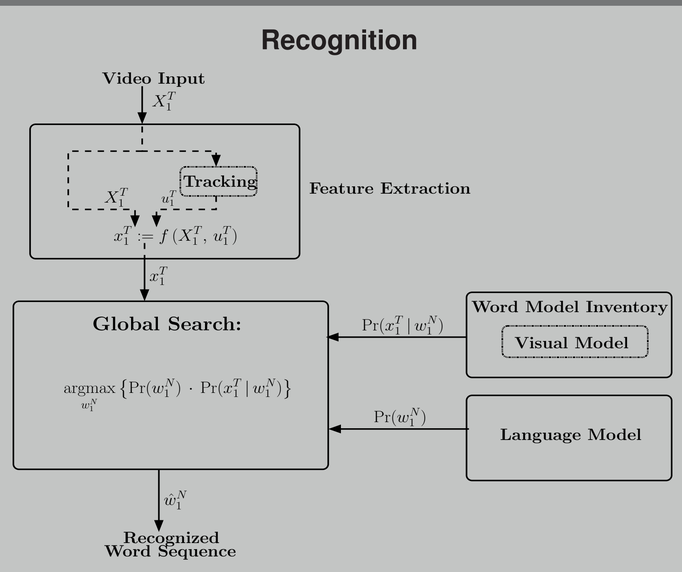
Yannick Gweth, Christian Plahl, and Hermann Ney

Human Language Technology and Pattern Recognition, RWTH Aachen University, Aachen, Germany

Introduction

- Automatic continuous sign language recognition
- Information conveyed visually, mainly using hands but also torso and facial expression
- Several speakers, appearence-based approach, no special hardware required
- Tracking the hands is a challenging task
  - Hands moving fast
  - High degree of freedom
- Tracking using RWTH dynamic programming tracker [Dreuw et al. 2006]
- Extraction of features describing the manual parameters
  - Appearence-based handpatch features
  - Neural network-based features (MLP)
- State-of-the-art results on the SIGNUM database using MLP-based features

Automatic Statistical Recognition



Goal:

- Find the gloss sequence which best explains input video stream
- Implicit sentences segmentation

Requirements:

- Gloss annotated video sequences

System Overview

Visual Modeling

- Related to the acoustic model in ASR
- HMM based, with separate Gaussian HMMs, globally pooled diag. cov. matrix
- Whole-word models

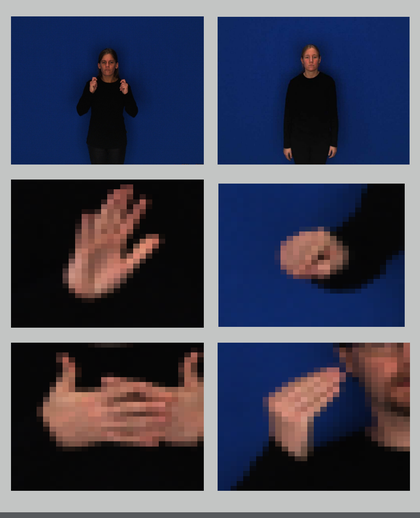
Language Modeling

- According to ASR: language model should have a greater weight than the visual model
- Trigram language model using the `SRILM` toolkit

Appearence-based image Features

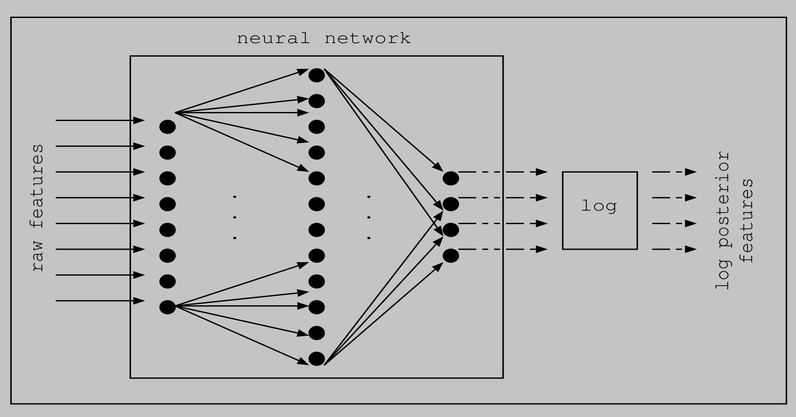
Features

- Appearance-based image features:
  - Thumbnails of video sequence frames (intensity images scaled to 32x32 pixels)
- Manual features:
  - Tracking: hand patch features
- Feature selection:
  - Concatenation of appearance-based and manual features
  - Sliding window for context modeling
  - Dimensionality reduction by PCA



MLP-based Features

MLP Feature extraction



- Input: PCA reduced appearence-based hand patches, sliding window
- 455 glosses + silence as targets
- One hidden layer with 1500 nodes
- Softmax activation in output layer
- Supervised training

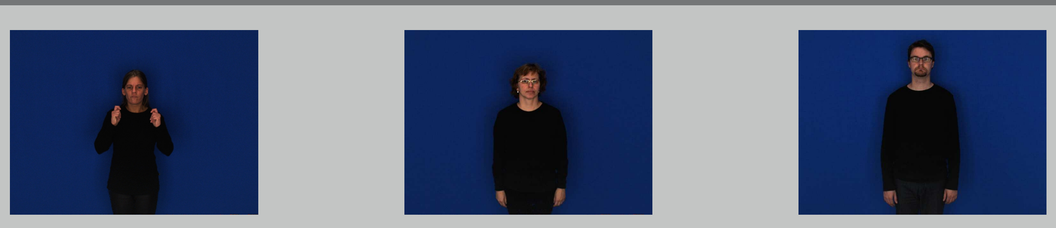
Requirements:

- initial alignment required (HMM system)

Synchrone Combination without Retraining

- Sign language is multimodal
- Differents models for recognition
- Modeled each modality separately
- Feature models weighted separately during recognition

SIGNUM database



- System evaluation on the SIGNUM database
- German Sign Language (DGS)
- 25 native signers
- Signers wear dark clothes in front of artificial blue background
- 780 continuous sentences
- One frontal camera
- Image resolution: 776 $\times$ 578 pixels
- 30 frame sper second

SIGNUM Signer Dependent Setup

Signer Dependent Setup

	Training	Test
# signers	1	1
# frames	46,638	6,751
# sentences	1809	531
# running words	11,109	2802
vocabulary size	455	-
OOVs (running)	-	1%

- One signer recorded the 780 sentences three times
- 3 $\times$ 603 for training and 3 $\times$ 177 for evaluation
- Perplexity (3-gram language model): Test 17.9, Training 97.5

Signer Dependent Results

MLP-based

Features	win	del / ins [%]	WER [%]
MLP	$\pm$1	2.0 / 1.1	13.0
	$\pm$2	1.7 / 2.6	14.7
	$\pm$3	2.1 / 3.8	17.4

Hand patch

Features	context	del / ins[%]	WER [%]
Hand patches	$\pm$1	4.6 /1.3	16.6
	$\pm$2	2.2 / 3.2	16.0
	$\pm$3	5.5 / 2.1	20.8

- Influence of concatenation of several features to a feature vector
- Using context information improves the recognition results

Combination

Features	del / ins [%]	WER [%]
Full image (F1)	7.2 / 2.3	28.2
Handpatches (F2)	2.2 / 3.2	16.0
MLP-based (F3)	2.0 / 1.1	13.0
F1 + F2 + F3	2.1 / 1.5	11.9

- Synchrone combination of the three models leads to 7% relative improvement compared to state-of-the-art[Von Agris et al. 2008]

Conclusions

- Significant improvements achieved using MLP-based features
- Quality of initial alignment to train the MLP impacts quality of generated features
- MLP generates feature distributions easily modeled by GHMM
- Sliding window on MLP input improves frame accuracy but not recognition results
- 19 % relative improvement compared to best appearence-based feature

Aknowledgement

This work received funding from the European Community's Seventh Framework Programme under grant agreement number 231424 (FP7-ICT-2007-3).

References

[Von Agris et al. 2007] U. von Agris and K. F. Kraiss. 2007. Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition In *Proceedings of GW2007-7th International Workshop on Gesture in Human-Computer Interaction and Simulation 2007–POSTER SESSION*, pages 10–11, Lisbon, Portugal.

[Dreuw et al. 2006] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. 2006. Tracking using dynamic programming for appearance-based sign language recognition. In *7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006.* , pages 293–298, Southampton, UK, April.

[Hermansky et al. 2000] H. Hermansky, D. Hellis and S. Sharma. 2000. Tandem connectionist feature stream extraction for conventional HMM systems. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000.*, pages 1635–1638, Istanbul, Turkey, June.

[Von Agris et al. 2008] U. von Agris, M. Knorr and K. F.. Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *8th IEEE International Conference on Automatic Face & Gesture Recognition, 2008..*, pages 1–6, Amsterdam, The netherlands, September.