Visual Modeling and Tracking Adaptation for Automatic Sign Language Recognition

Philippe Dreuw – Joint work with Thomas Deselaers, and Hermann Ney

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

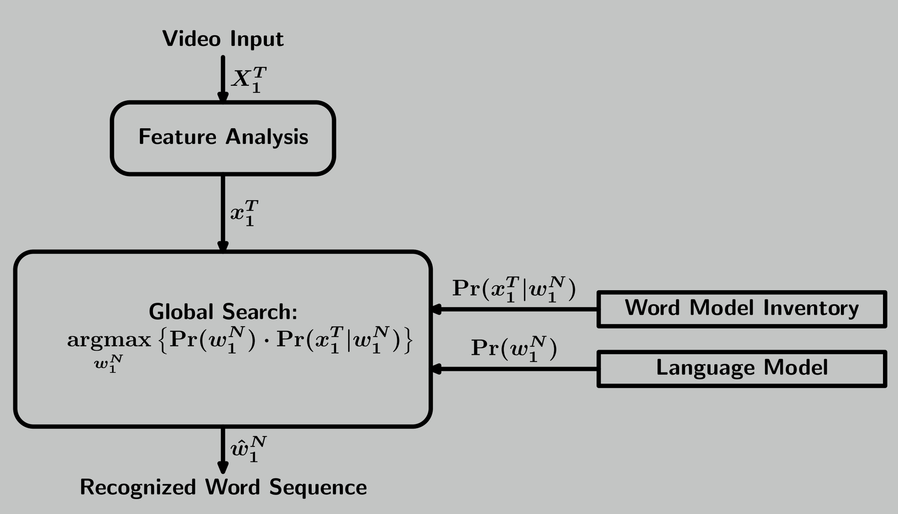
RWTH AACHEN UNIVERSITY

Introduction

- automatic sign language recognition system
- necessary for communication between deaf and hearing people
- continuous sign language recognition, several speakers, vision-based approach, no special hardware
- large vocabulary speech recognition (LVSR) system to obtain a textual representation of the signed sentences
- evaluation of speech recognition techniques on publicly available sign language corpus

Automatic Sign Language Recognition (ASLR)

- differences to speech recognition: simultaneousness
- similar to speech recognition: temporal sequences of images
- important features
  - hand-shapes, facial expressions, lip-patterns
  - orientation and movement of the hands, arms or body
- HMMs are used to compensate time and amplitude variations of the signers
- goal: find the model which best expresses the observation sequence



System Overview

Visual Modeling (VM)

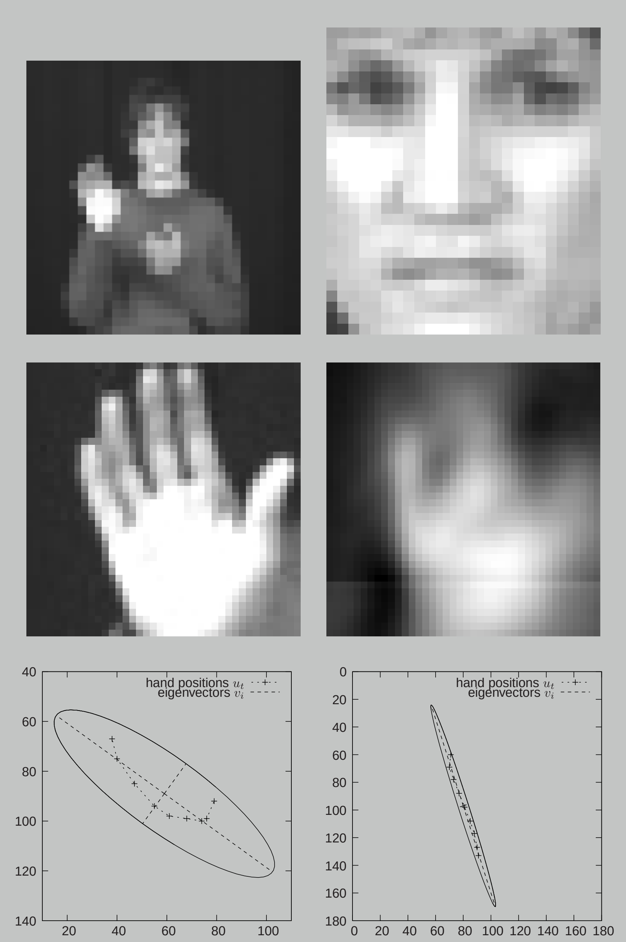
- related to the acoustic model in ASR
- HMM based, with separate GMMs, globally pooled diag. cov. matrix
- monophone whole-word models
- pronunciation handling

Language Modeling (LM)

- according to ASR: LM should have a greater weight than the VM
- trigram LM using the SRILM toolkit, with modified Kneser-Ney discounting with interpolation

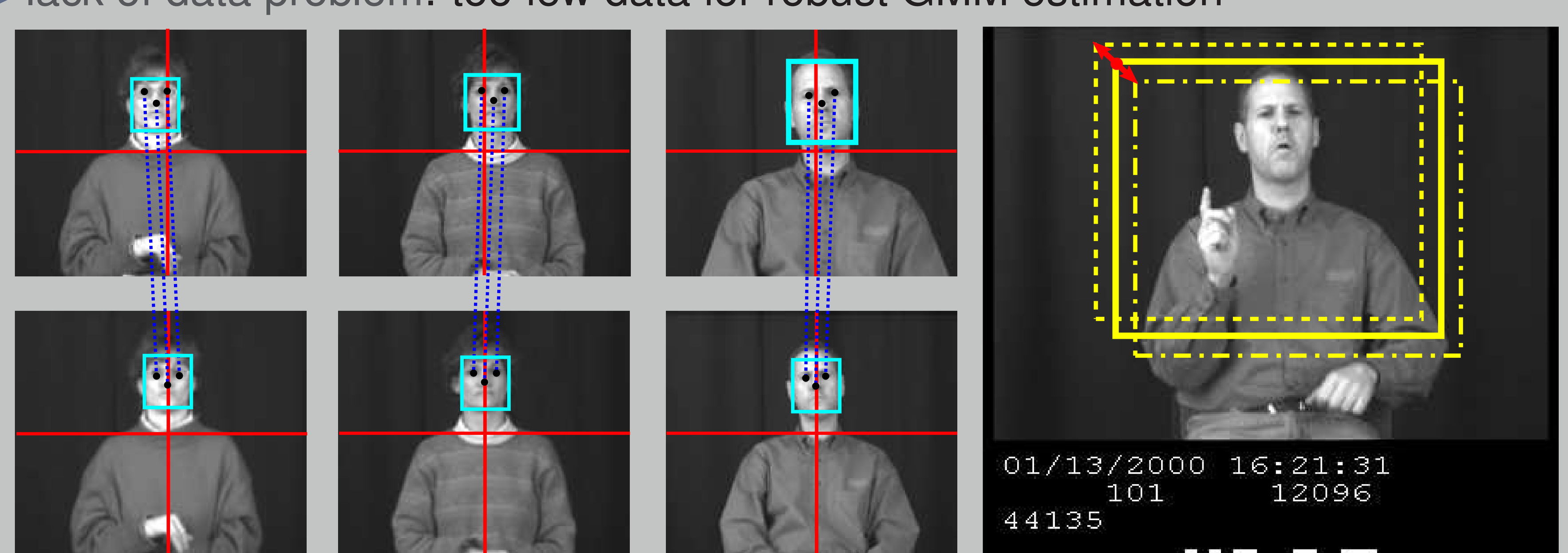
Features

- appearance-based image features: for baseline system
  - thumbnails of video sequence frames (intensity images scaled to 32x32 pixels)
  - give a global description of all (manual and non-manual) features proposed in linguistic research
- manual features:
  - tracking: hand position, hand velocity, and hand trajectory features
- feature selection:
  - concatenation of appearance-based and manual features
  - sliding window for context modeling
  - dimensionality reduction by PCA and/or LDA



Visual Speaker Alignment (VSA) and Virtual Training Samples (VTS)

- visually align speakers: extract scale and speaker independent features
- lack of data problem: too few data for robust GMM estimation

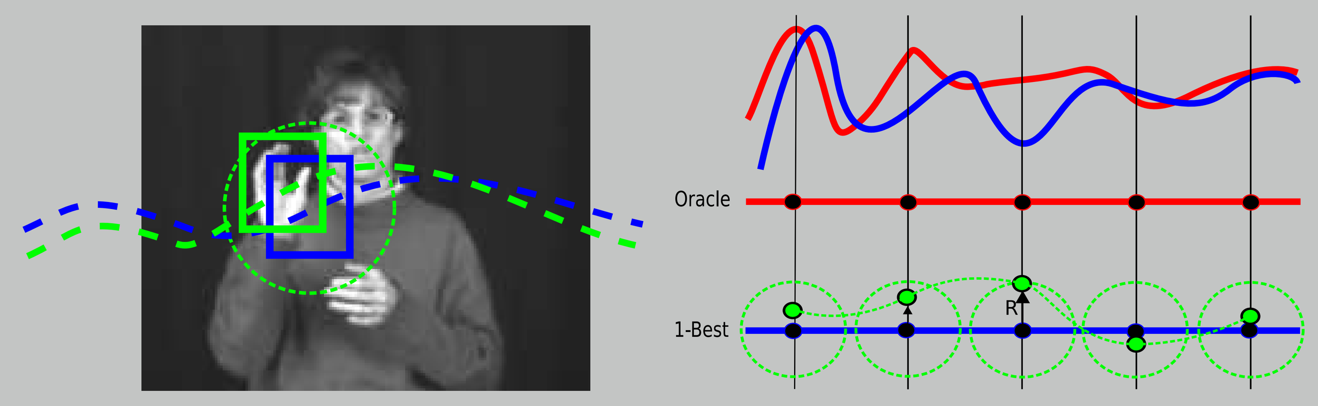


Feature Adaptation and System Combination

Feature Adaptation

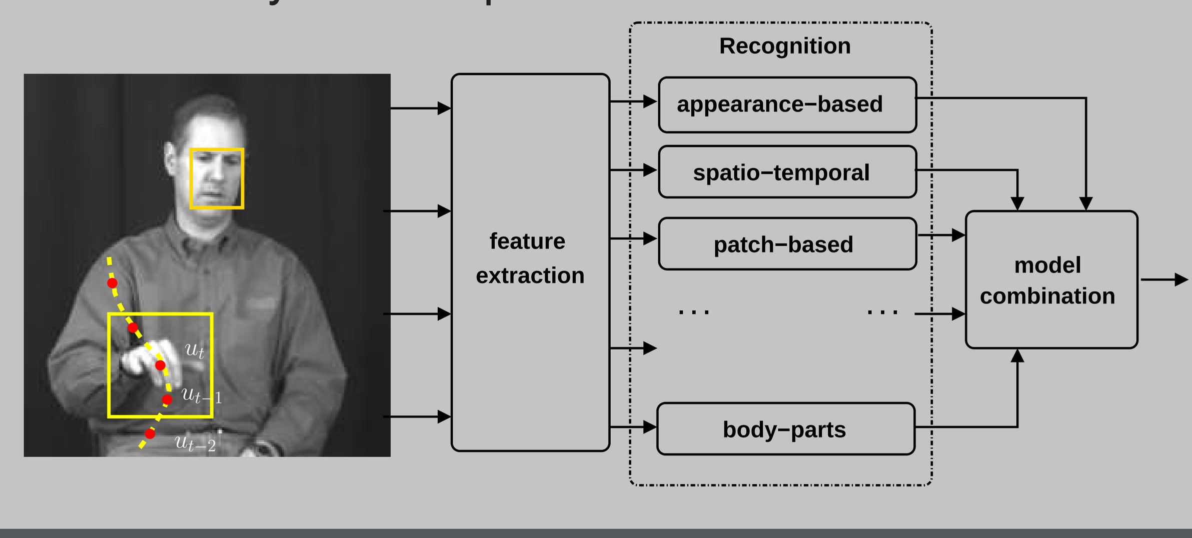
- problem: tracking as preprocessing, optimized only w.r.t. motion
- model-based tracking path adaptation: consider locations around given tracking path u_1^T within range R
- features are adapted during recognition w.r.t. hypothesized word sequence:
- VM probability changes as follows:

$$\Pr(x_1^T, s_1^T|w_1^N) = \prod_{t=1}^{T} \left\{ \max_{\substack{\delta \in \{(x,y): \\ -R \leq x,y \leq R\}}} \{p(\delta) \cdot p(f(X_t, u_t + \delta)|s_t, w_1^N)\} \cdot p(s_t|s_{t-1}, w_1^N) \right\}$$



Model and System Combination

- log-linear combination of independently trained models
- profit from independent alignments (e.g. performing well for long and short words)
- profit from different feature extraction approaches
- ROVER over different system outputs and confidences



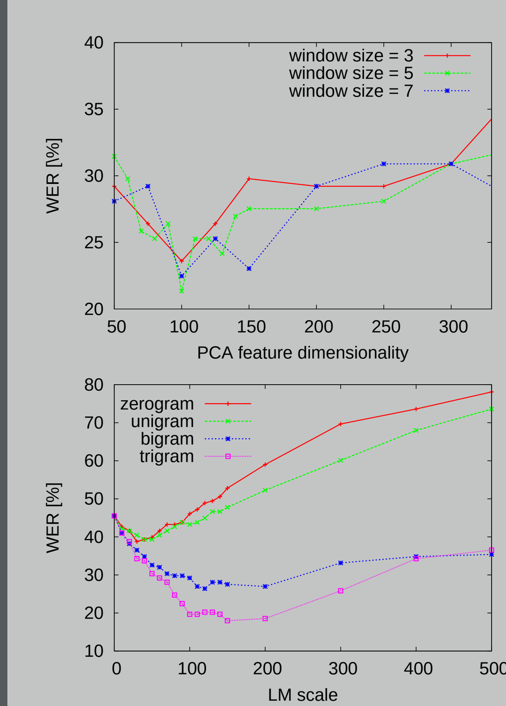
Experimental Results

Database

- system evaluation on the RWTH-BOSTON-104 database
  - 201 sentences (161 training and 40 test)
  - vocabulary size of 104 words
  - 3 speakers (2 female, 1 male)
  - corpus is annotated in glosses
  - 26% of the training data are singletons

Results

- Baseline System



Features / Adaptation	WER[%]			
	Baseline	VSA	VTS	VSA+VTS
Frame 32×32	35.62	33.15	27.53	**24.72**
PCA-Frame (200)	30.34	27.53	19.10	**17.98**
Hand (32×32)	45.51	33.15	20.79	21.91
+ distortion ($R = 10$)	41.03	29.78	16.29	16.85
+ δ-penalty	35.96	26.40	**15.73**	16.85
PCA-Hand (70)	44.94	34.27	63.48	20.22
+ distortion ($R = 10$)	56.74	34.83	28.08	15.73
+ δ-penalty	32.58	24.16	25.84	**14.04**

- ROVER (4 systems): **12.9% WER**

Conclusion

- LVSR system is suitable for vision-based continuous sign language recognition
- many of the principles known from ASR can directly be transfered
- important for ASLR: temporal contexts, pronunciation handling, language modelling, and model combination
- VSA and VTS effects are cumulative, can be applied to any vision-based approach
- outlook: connection of recognizer output to a statistical machine translation system achieved promising translation results