Efficient Approximations to Model-based Joint Tracking and Recognition of Continuous Sign Language

Philippe Dreuw, Jens Forster, Thomas Deselaers, and Hermann Ney

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

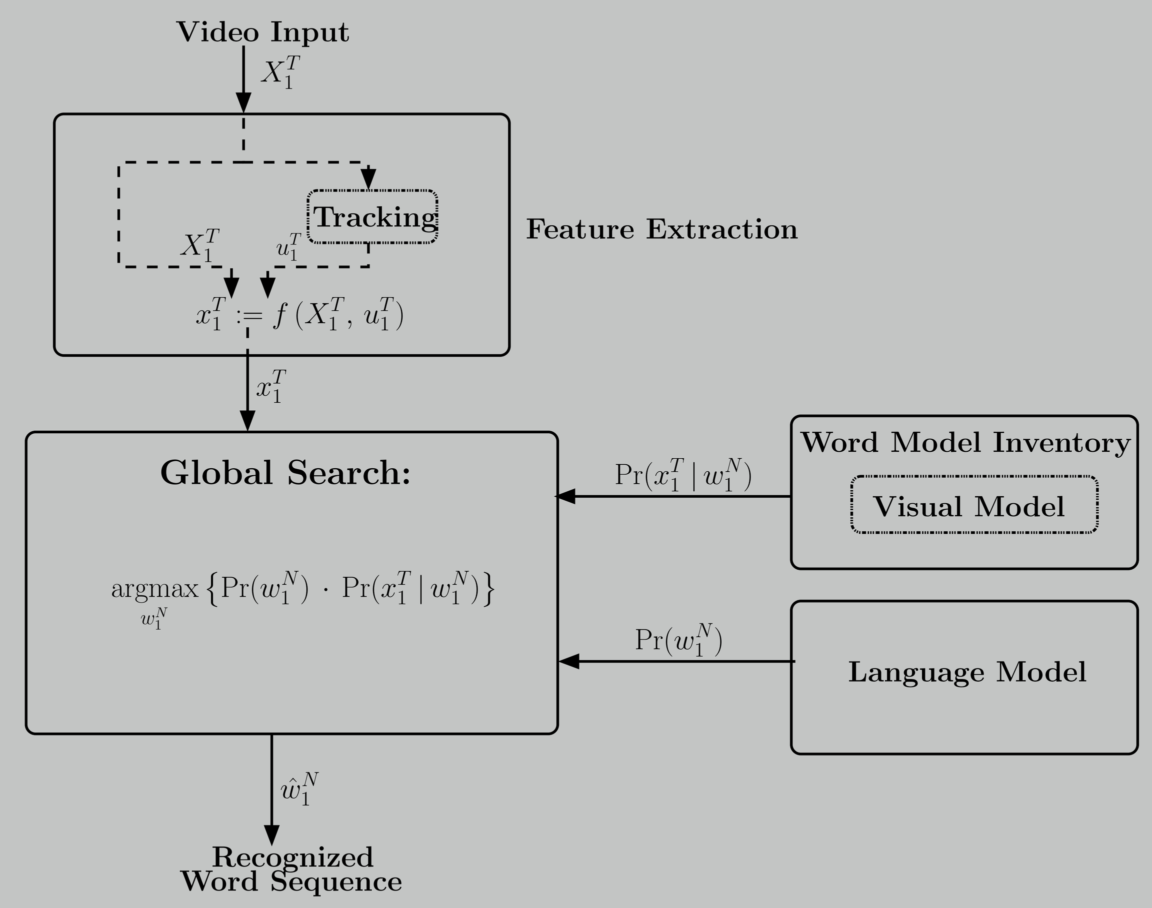
RWTH AACHEN UNIVERSITY

Introduction

- automatic continuous sign language recognition system
- necessary for communication between deaf and hearing people
- problems
  - lack of data
  - early tracking decisions lead to recognition errors

Automatic Sign Language Recognition (ASLR)

- goal: find the word sequence which best expresses the observation sequence (i.e. the tracked features)



- Bayes' decision rule:

$$x_1^T \longrightarrow r(x_1^T) = \underset{w_1^N}{\operatorname{argmax}} \left\{ \Pr(w_1^N) \cdot \Pr(x_1^T | w_1^N) \right\} \quad (1)$$

$$\hat{w}_1^N = \underset{w_1^N}{\operatorname{argmax}} \left\{ p(w_1^N) \max_{s_1^T} \prod_{t=1}^{T} \left\{ p(f(X_t, u_t) | s_t, w_1^N) \cdot p(s_t | s_{t-1}, w_1^N) \right\} \right\}$$

- problem: early tracking decisions in preprocessing steps

System Overview

Visual Modeling

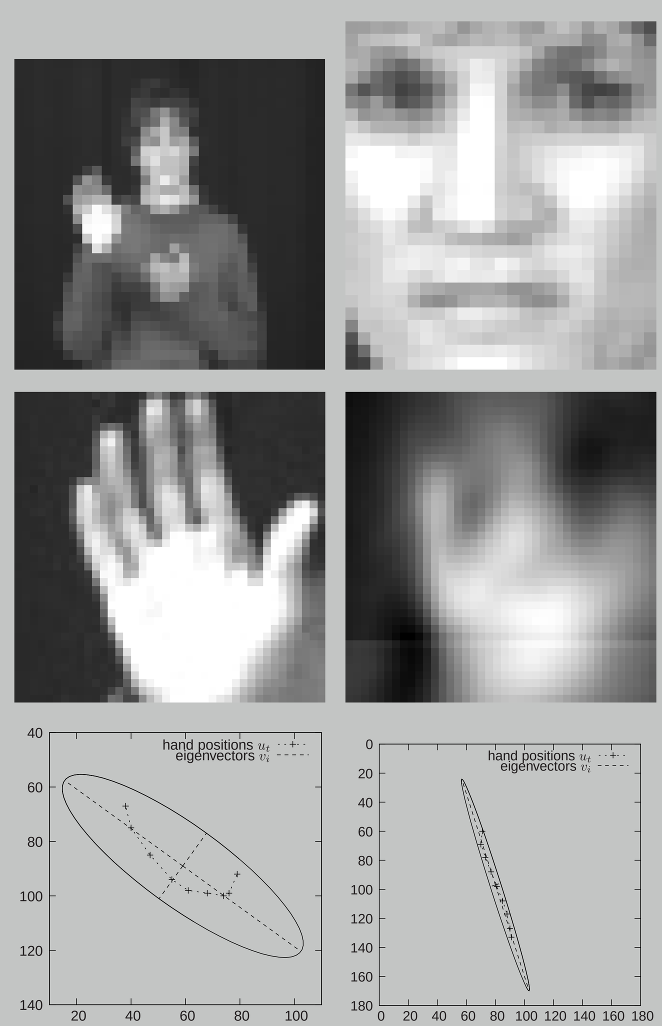
- related to the acoustic model in ASR
- HMM based, with separate GMMs, globally pooled diag. cov. matrix
- monophone whole-word models
- pronunciation handling

Language Modeling

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

Features

- appearance-based image features:
  - thumbnails of video sequence frames (intensity images scaled to 32x32 pixels)
- manual features:
  - tracking: hand trajectory features
- feature selection:
  - concatenation of appearance-based and manual features
  - sliding window for context modeling
  - dimensionality reduction by PCA and/or LDA



Virtual Training Samples (VTS)

- lack of data problem: too few data for robust GMM estimation
- here: several region-of-interests (ROI) are cropped from the original video at each frame
- ROI cropping center (x, y) is shifted by δ pixels in x- and y-direction
- example: for $\delta = \pm 1$, the training corpus is already enlarged by a factor of nine.

Hand Tracking

- optimize the tracking decision considering the full sequence using dynamic programming (DP) (see [Dreuw et. al FG2006])
- The DP tracking consists of two steps
  1. obtain scores D and backpointers B

$$D(t, x, y) = \max_{x', y' \in M(x,y)} \{(D(t-1, x', y') - \mathcal{J}(x', y', x, y)\} + d(x', y', x, y, X_{t-1}^t) \quad (2)$$

$$B(t, x, y) = \underset{x', y' \in M(x,y)}{\operatorname{argmax}} \{(D(t-1, x', y') - \mathcal{J}(x', y', x, y)\}$$

  2. traceback process reconstructs the best path $t \rightarrow u_t = (x, y)$ using the score table D and the backpointer table B starting from time step T

$$u_{t-1} = B(t, u_t) \quad \text{with} \quad u_T = \underset{(x,y)}{\operatorname{argmax}} \{D(T, x, y)\} \quad (3)$$

- allows to optimze tracking decisions over full temporal context
- problem: early tracking decisions in preprocessing, optimized only w.r.t. motion, etc.

Integrated Tracking and Recognition

- postpone the tracking decisions to the end of the recognition phase
- simultaneous optimization: tracking positions u_1^T optimal w.r.t. a tracking criterion **and** a hypothesized word sequence w_1^N

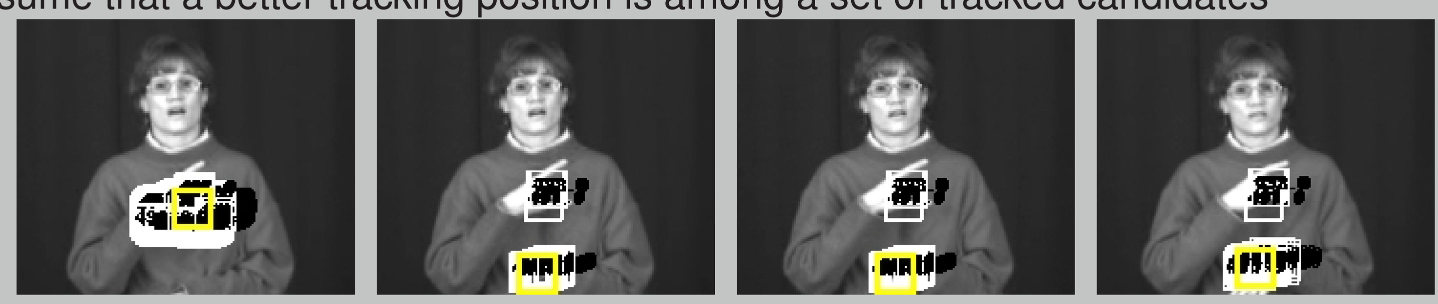
$$\Pr(x_1^T | w_1^N) = \sum_{[s_1^T]} \sum_{[u_1^T]} \Pr(X_1^T, s_1^T, u_1^T | w_1^N)$$

$$\propto \max_{[s_1^T]} \max_{[u_1^T]} \prod_{t=1}^{T} \Big[\underbrace{\Pr(X_t | s_t, u_t, w_1^N)}_{\text{emission}} \cdot \underbrace{\Pr(s_t | s_{t-1})}_{\text{state transition}} \cdot \underbrace{\Pr(u_t | u_{t-1}, X_{t-1}^t)}_{\text{position transition}} \Big] \quad (4)$$

- problem: very high time and memory complexity

Efficient Approximations: Rescoring and Feature Adaptation

- $\Pr(f(X_t, u_t) | s_t, w_1^N)$ depends on the quality of the hand tracking position u_t
- we assume that a better tracking position is among a set of tracked candidates



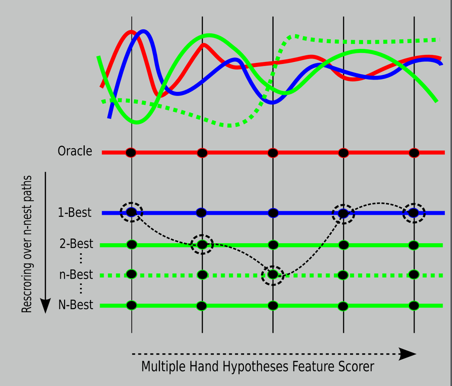
n-Best Tracking List Rescoring

- by tracing back multiple times over the sorted score table D and the backpointer table B.
- Eq. (3) changes for $i = 1, .., n$ as follows:

$$u_{t-1,i} = B(t, u_{t,i}) \quad \text{with} \quad u_{T,i} = \underset{\substack{(x,y) \notin \\ \{u_{T,1}, \ldots, u_{T,i-1}\}}}{\operatorname{argmax}} \; D(T, x, y)$$

- Visual Model in Eq. (1) changes as follows:

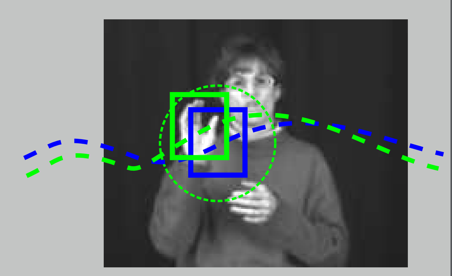
$$\Pr(x_1^T, s_1^T | w_1^N) = \prod_{t=1}^{T} \left\{ \max_{i: u_1^T := (u_{1,i}, \ldots, u_{T,i})} \left\{ p(f(X_t, u_t) | s_t, w_1^N) \right\} \cdot p(s_t | s_{t-1}, w_1^N) \right\}$$



Multiple Hands Hypotheses

- consider at each time step t a set of n possible hand positions $\{u_{t,1}, .., u_{t,n}\}$
- Visual Model in Eq. (1) changes as follows:

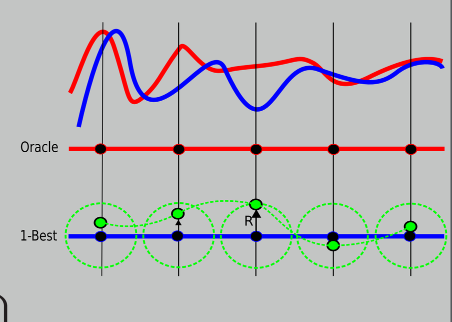
$$\Pr(x_1^T, s_1^T | w_1^N) = \prod_{t=1}^{T} \left\{ \max_{i=1,\ldots,n} \left\{ p(f(X_t, u_{t,i}) | s_t, w_1^N) \right\} \cdot p(s_t | s_{t-1}, w_1^N) \right\}$$



Path Distortion Model

- consider positions around given tracking path u_1^T within range R
- Visual Model in Eq. (1) changes as follows:

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



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	DEL	INS	SUB	errors	WER %
Frame (32x32)	43	6	16	65	35.62
PCA-Frame (200)	40	9	18	27	30.34
Hand (32x32)	31	7	43	81	45.51
PCA-Hand (70)	40	10	21	49	44.94

- n-best list rescoring and multiple hand hypotheses (MHH)

Delay Δ	WER[%]					
	$M = \pm 1$			$M = \pm 10$		
	1-best	n-best	MHH	1-best	n-best	MHH
Full	80.34	76.97	76.40	**45.51**	**45.51**	**45.51**
100	79.78	75.28	73.03	45.51	45.51	45.51
25	70.79	64.61	66.29	56.18	50.56	53.37
10	69.10	67.98	65.17	63.48	60.11	58.99
1	91.01	83.71	65.17	91.01	83.71	65.17

- Path Distortion Model

Features / Rescoring	WER[%]			
	pixel values		PCA transformed	
	Baseline	VTS	Baseline	VTS
Frame 32×32	35.62	**27.53**	30.34	**19.10**
Hand (32×32)	45.51	20.79	44.94	15.73
+ distortion ($R = 10$)	41.03	16.29	56.74	12.92
+ δ-penalty	35.96	**15.73**	32.58	**11.24**

<surname>@cs.rwth-aachen.de