Visual Modeling and Feature Adaptation in Sign Language Recognition

Philippe Dreuw and Hermann Ney
dreuw@cs.rwth-aachen.de

ITG 2008, Aachen, Germany – Oct 2008

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

1 Introduction

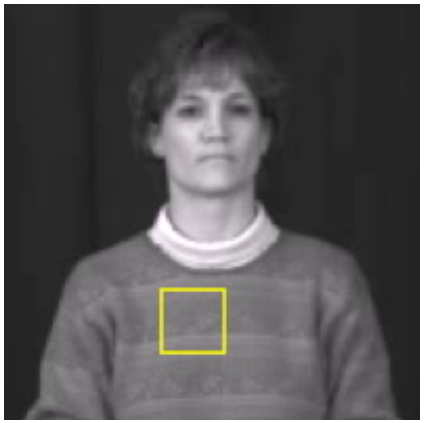
- **Hearing-to-Deaf communication:**
  - **only 30-40% of speech can be understood through lipreading**
  - **most accessibility aids assume strong literacy skills**
  - **television closed captioning may exceed reading level**
- **Deaf-to-Hearing communication:**
  - **most people of the Deaf community have poor to moderate writing skills**
  - **live interpreters are not always available**

Key idea:

- **a sequence of gestures is transcribed as a sequence of GLOSSES.**
- **GLOSSES are then translated into a spoken target language (e.g. English).**

Application: Sign-to-Speech

Recognition: Sign-to-Text (Video →Glosses)



⇓

Translation: Text-to-Text (Glosses →Text)

JOHN FISH WONT EAT BUT CAN EAT CHICKEN

John will not eat fish but eats chicken

⇓

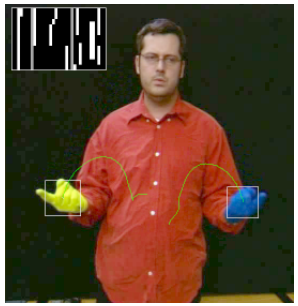
Synthesis: Text-to-Speech (Text →Audio)

021.wav

2 Automatic Sign Language Recognition

- what features do we need?
  - manual components: hand motion / form / orientation / location
  - non-manual components: mimic, eye gaze, body / head orientation

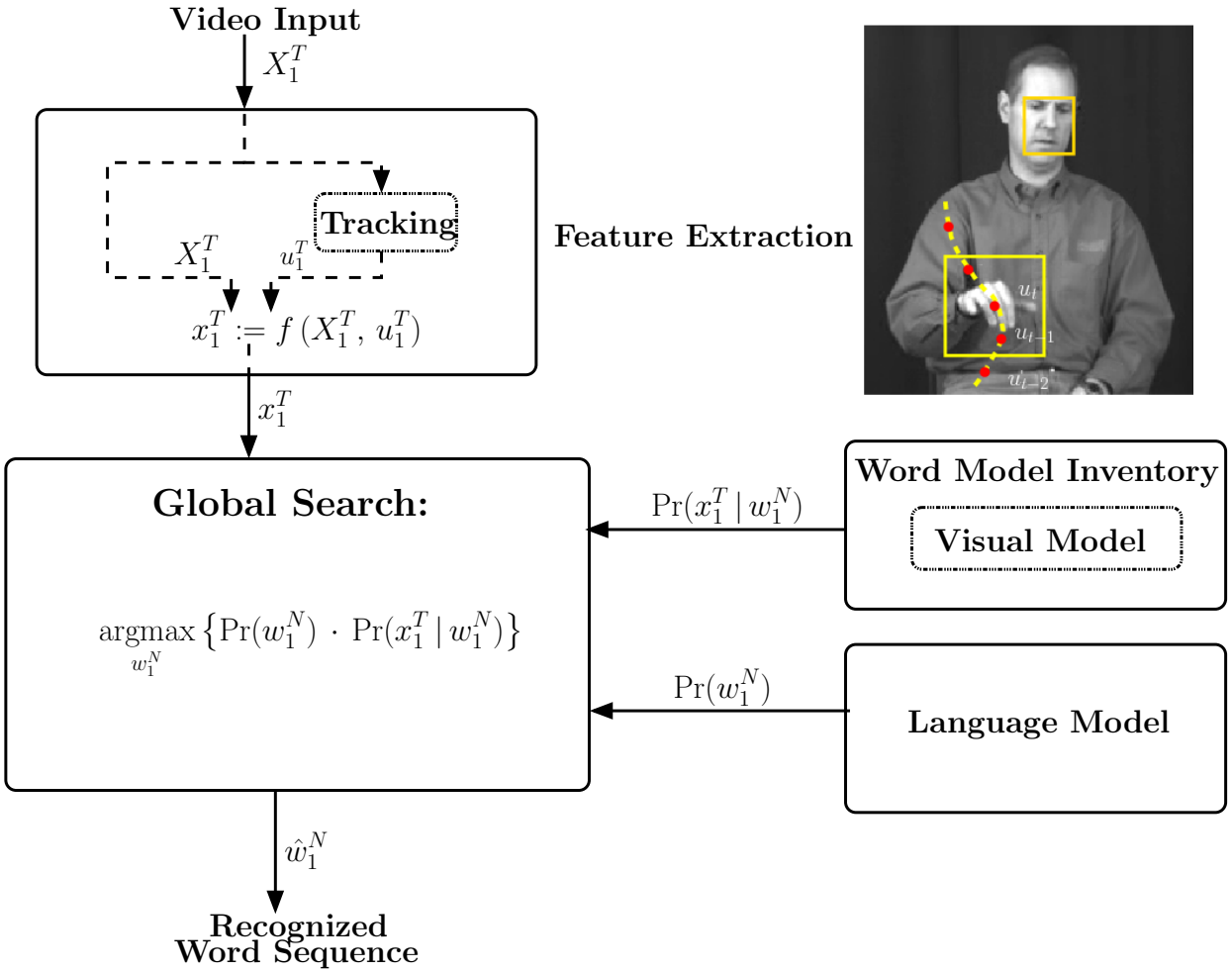
- different approaches / assumptions
  - special hardware
  - computer vision


→ only the vision-based approaches do not restrict the way of signing

→ our approach/setup: similar to speech recognition

System Overview



$$\hat{\boldsymbol{w}}_1^{\boldsymbol{N}} = \underset{\boldsymbol{w}_1^{\boldsymbol{N}}}{\operatorname{argmax}} \left\{ \boldsymbol{p}(\boldsymbol{w}_1^{\boldsymbol{N}}) \max_{\boldsymbol{s}_1^{\boldsymbol{T}}} \prod_{\boldsymbol{t}=1}^{\boldsymbol{T}} \left\{ \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{X}_t, \boldsymbol{u}_t) | \boldsymbol{s}_t, \boldsymbol{w}_1^{\boldsymbol{N}}) \cdot \boldsymbol{p}(\boldsymbol{s}_t | \boldsymbol{s}_{t-1}, \boldsymbol{w}_1^{\boldsymbol{N}}) \right\} \right\} \quad (1)$$

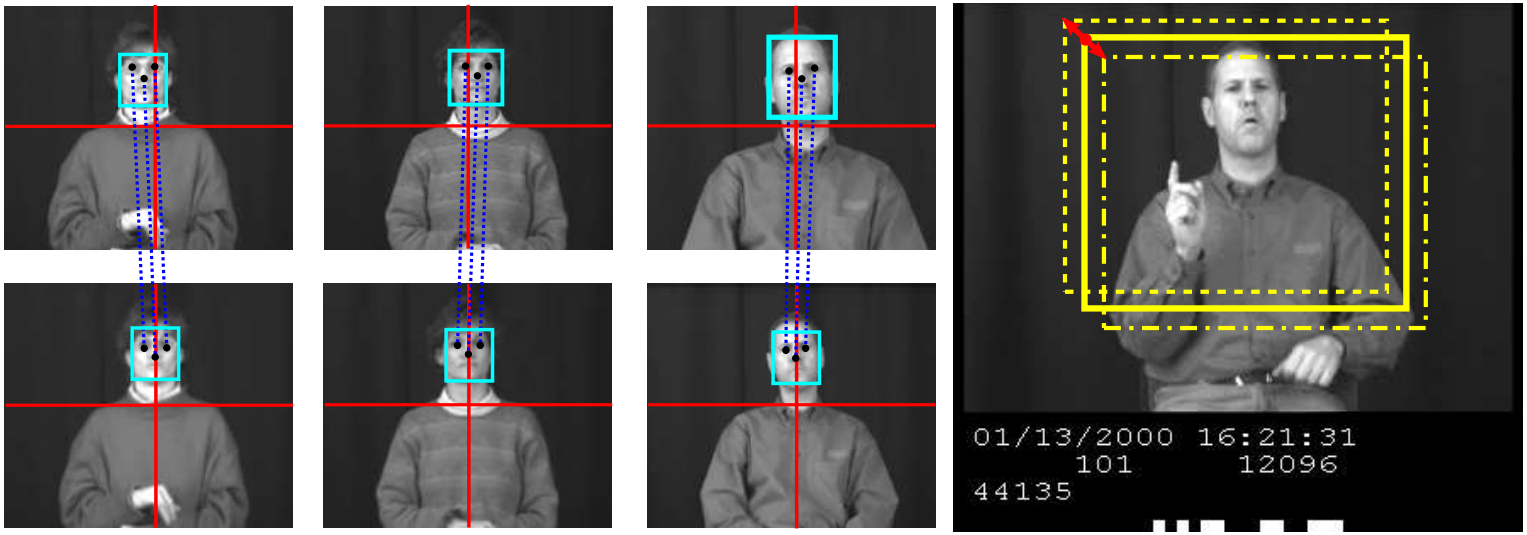
3 Visual Speaker Alignment and Virtual Training Data

- **2 Problems:**
  - **appearance-based features are position and scale dependent**
  - **lack of training data in sign language processing**

- **Ideas:**

  (a) automatic speaker alignment based on face detection

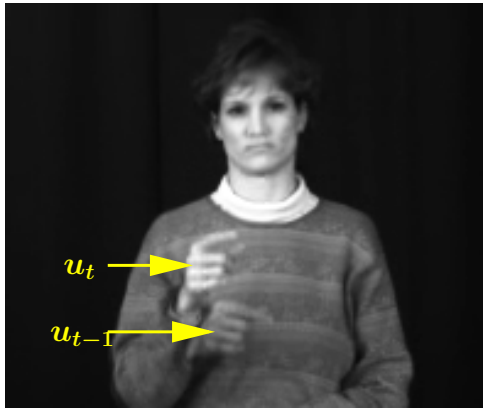
  (b) virtual training samples generation



(a) (b)

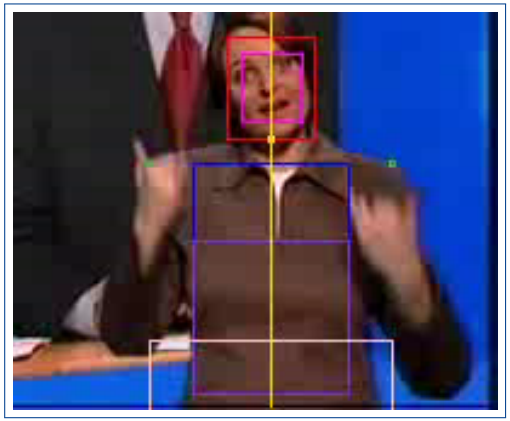
4 Tracking Shortcomings

- **tracking as pre-processing:**
  - **path only optimized w.r.t. a tracking criterion (e.g. motion, color, etc.)**
  - **early tracking decisions can lead to recognition errors**
- **examples**



overlay of two consecutive frames with labeled hand positions

head and hand tracking on RWTH-Boston-104 database

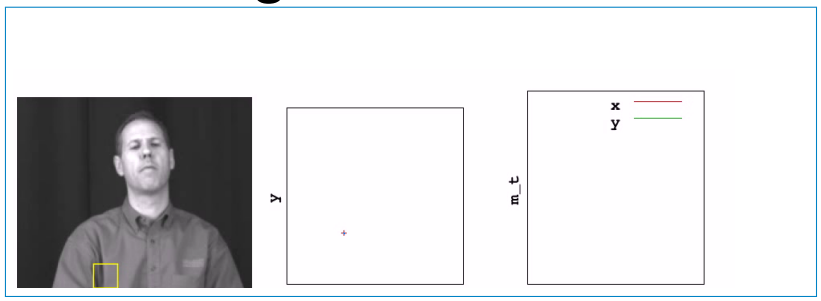


head and hand tracking on RWTH-Phoenix database with spatial body pose model

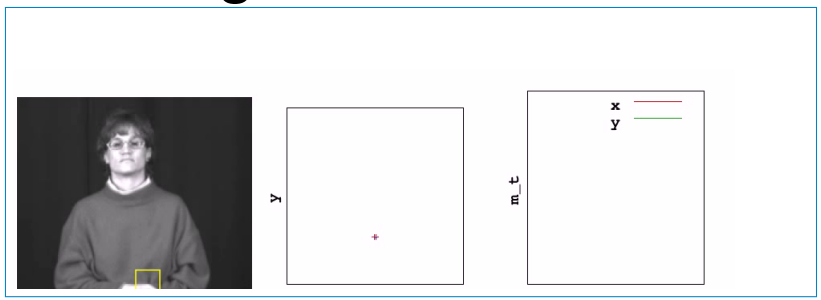
head tracking with Viola & Jones method

Tracking Shortcomings

- hand tracking confusion w/o recognition errors



- hand tracking confusion w/ recognition errors



- without model information:
  - problematic recovery from errors
  - no differentiation between similar objects

Tracking Shortcomings

- **Motivation:**
  we expect that a (locally) better tracking position is among a set of candidates




- **Idea: Joint Tracking and Recognition**

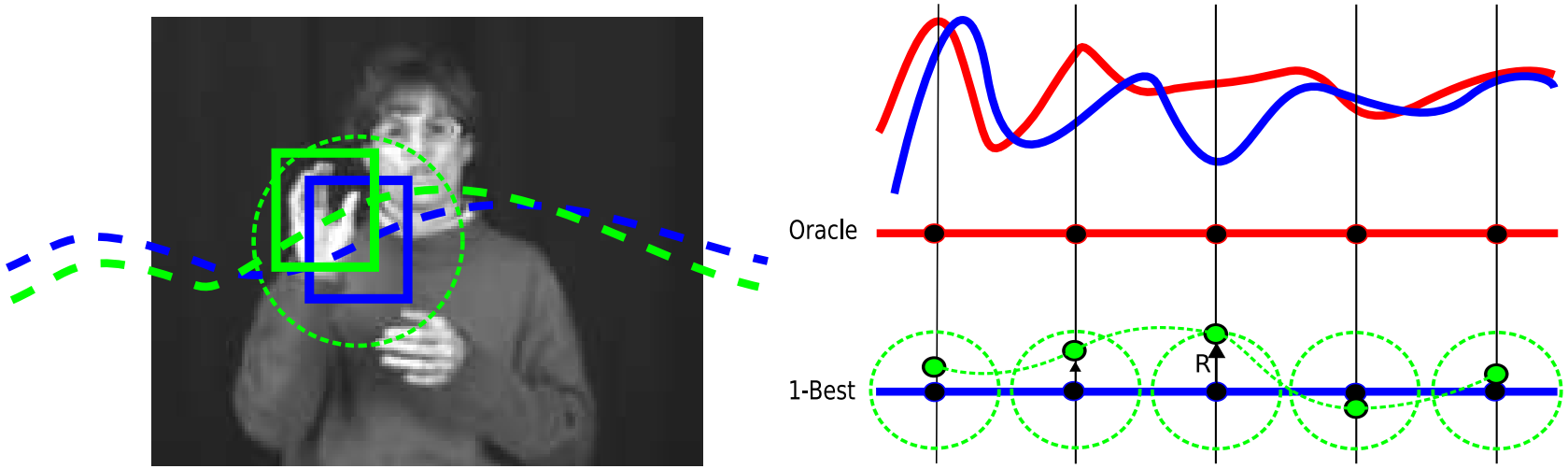
  $\longrightarrow$ **simultaneous optimization of tracking path u_1^T**
  and hypothesized word sequence w_1^N (late fusion)

- **Problem:**
  - **full search has high time and memory complexity**

  $\longrightarrow$ **approximation: model-based tracking path adaptation**

Approximation: Model-Based Tracking Path Adaptation

- consider positions around given tracking path u_1^T within range R



- visual model in Eq. (1) changes as follows:

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

$$\text{with} \quad p(\delta) = \frac{\exp(-\delta^2)}{\exp(\sum_{\delta'} -\delta'^2)}.$$

5 Experimental Results

- **results using the path distortion model for visual speaker alignment (VSA) and virtual training samples (VTS)**

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
+ δ-penalty	35.96	26.40	**15.73**	16.85
PCA-Hand (70)	44.94	34.27	15.73	20.22
+ δ-penalty	32.58	24.16	14.61	**14.04**

Thank you for your attention

Philippe Dreuw

dreuw@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/