Continuous Sign Language Recognition
Approaches from Speech Recognition

Philippe Dreuw
dreuw@i6.informatik.rwth-aachen.de

Invited Talk at DCU – 9. June 2006

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

RTWH Aachen University - Computer Science Department 6

- Statistical Machine Translation (SMT)
  - ≈ 13 people
  - focus on phrase-based SMT
  - Arabic-English / Chinese-English / Spanish-English
    - 220M running words (parallel data), 600M words for LM (monolingual)
- Automatic Speech Recognition (ASR)
  - ≈ 10 people
  - focus on system combination, unsupervised training, speaker adaption, and open vocabulary speech recognition
  - Arabic, Chinese, English, Spanish
- Computer Vision (CV)
  - ≈ 3 people
  - main focus on image retrieval and object detection
    - image retrieval for medical applications
  - new focus on sign language recognition to combine the knowledge of ASR and CV

Outline

Introduction

Features & System Overview

Databases & Experimental Results

Outlook

Automatic Sign Language Recognition

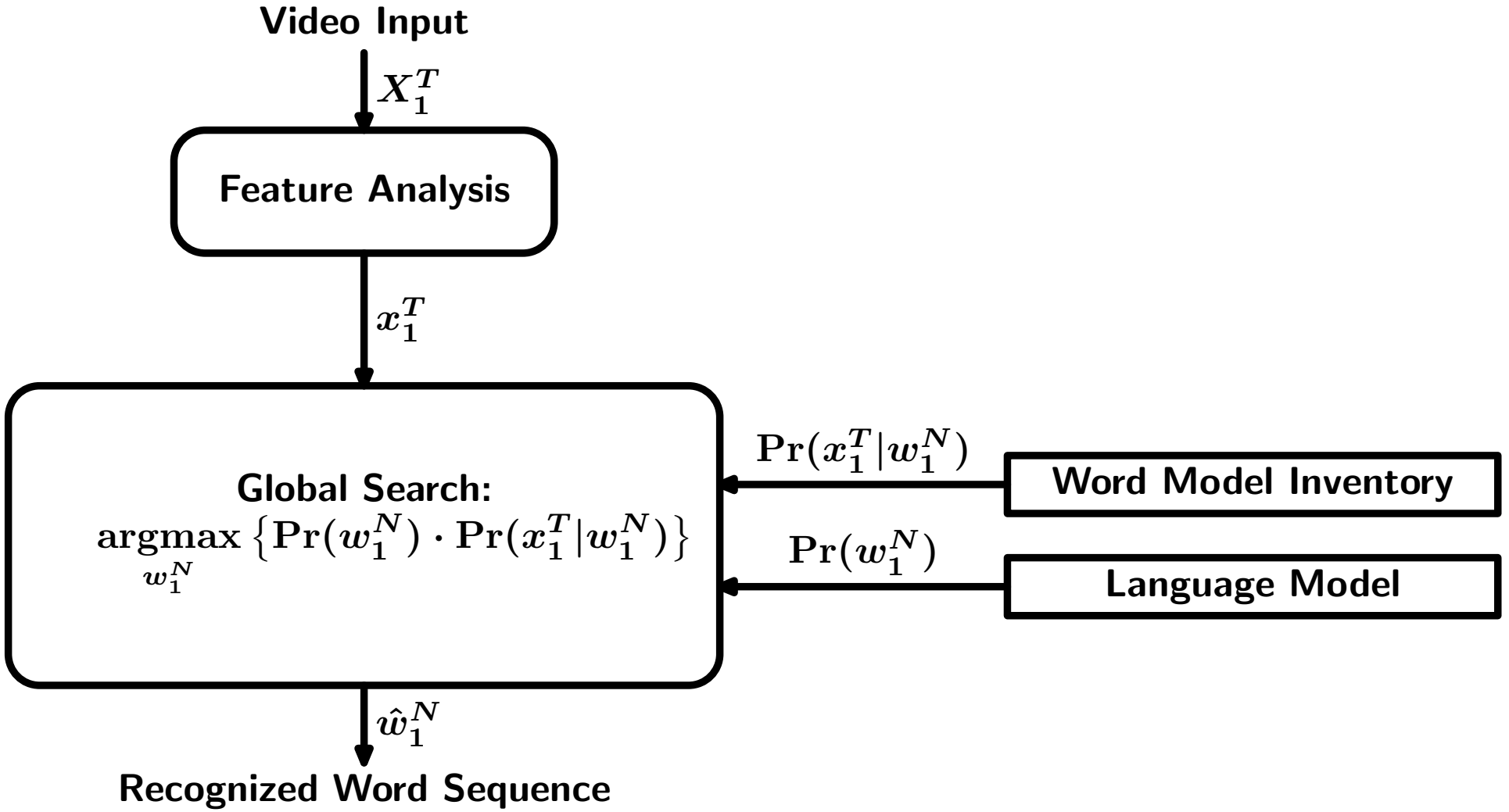
- Many topics are combined
- Capturing problems, tracking problems, segmentation problems
- Most systems: very person dependent, recognition of isolated signs
- Co-articulation effects in continuous sign language and dialects
- No publicly available corpora

- Our Approach/Setup: similar to speech recognition
  - Recognition of continuous sign language
  - Training with sentences (unknown word boundaries)
  - Person independent training and recognition
  - Data: videos showing the frontal view of the speakers
  - No colored gloves, markers or calibration

⇒use RWTH-i6 large vocabulary speech recognition system SPRINT

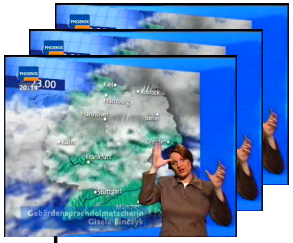
Automatic Sign Language Recognition

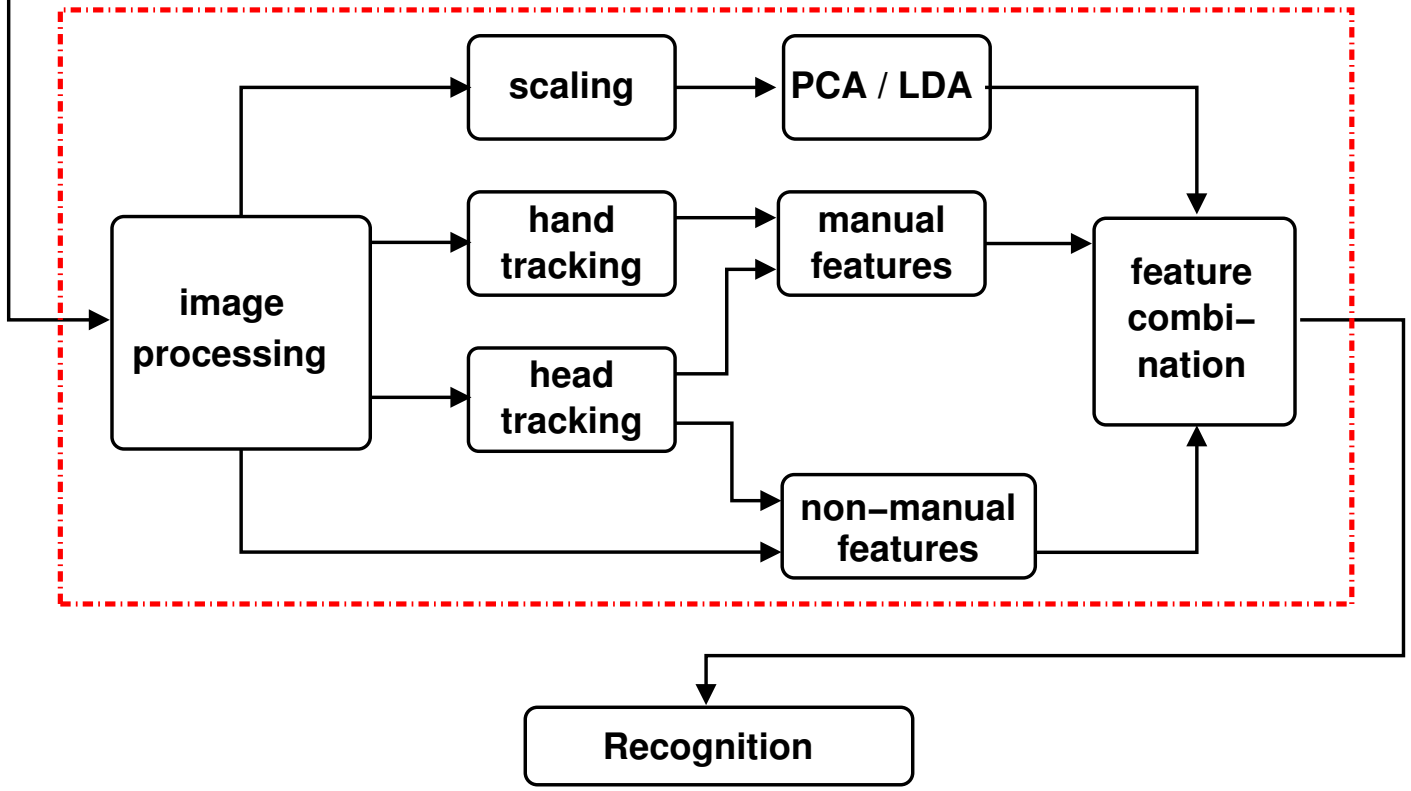
Bayes' Decision Rule



System Overview - Features

Feature analysis in FLOW = flexible signal analysis network





JETZT WETTER+VORAUS+SAGEN MORGEN FREITAG ACHTZEHN MAERZ

System Overview - Features

Baseline Setup:

- downscaled frame (32×32 pixel)
- Reduced to 110 components using PCA
- Trigram language model (using SRI LM Toolkit)

$$\Pr(w_1^N) = \prod_{n=1}^{N} p(w_n|w_{n-2}, w_{n-1})$$

- Important features for sign language recognition:
  - hand patch/position/motion/trajectory, ...
  - facial expression, ...
  - body pose, spatial features, ...

`http://www.speech.sri.com/projects/srilm/`

System Overview - Tracking

Tracking: Introduction

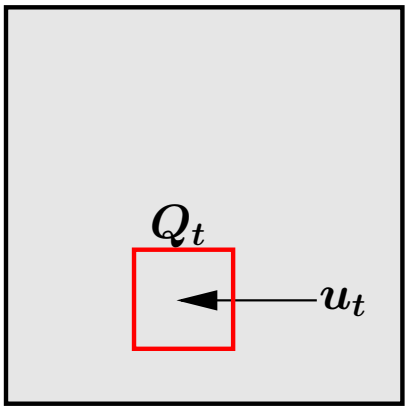
- Application: Tracking of dominant hand
- Goal: Find best path u_1^T of object positions in a sequence X_1^T of images.
- Avoid possibly wrong local decisions.
- Challenges:
  - hand in front of the face
  - temporary disappearance
  - crossing hands

Two Steps:

1. **Score calculation:** calculate a global score $S(t, x, y)$ and a backpointer table B for the best tracking until time step t which ends in position (x, y)
2. **Traceback:** reconstruct the best path $t \rightarrow (x, y)$ using S and B

System Overview - Tracking

$$
\begin{aligned}
u &:= (x, y) \\
Q &:= \left\{ (x, y) : \begin{array}{l} -w \leq x \leq w, \\ -h \leq y \leq h \end{array} \right\} \\
Q_t &:= \{ u_t + u : u \in Q \} \\
X'_t[u] &:= X_t[u] - X_{t-1}[u]
\end{aligned}
$$



X_t

X'_t

Example: Maximum Gradient - Maximize motion

$$
\operatorname{argmax} u_1^T \left\{ \sum_{t=1}^{T} \left(\sum_{u \in Q_t} \left(X'_t[u] \right)^2 + \sum_{u \in Q_{t-1}} \left(X'_t[u] \right)^2 - \lambda \| u_t - u_{t-1} \|^2 \right) \right\}
$$

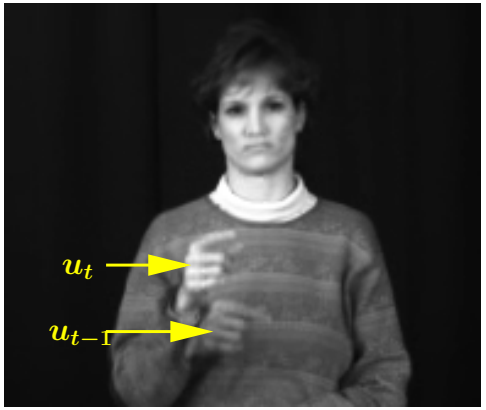
System Overview - Tracking

- **Application: Tracking of head**
- **Challenges:**
  - **hand in front of the face**
  - **head rotation, strong facial expressions**
  - **eigenfaces and skin color can only be used to produce candidate regions**
- **Idea: combination of skin color score s_c and Eigenfaces s_f for head tracking**

$$s(t,x,y) = (1-w) \cdot s_c(t,x,y) + w \cdot s_f(t,x,y)$$

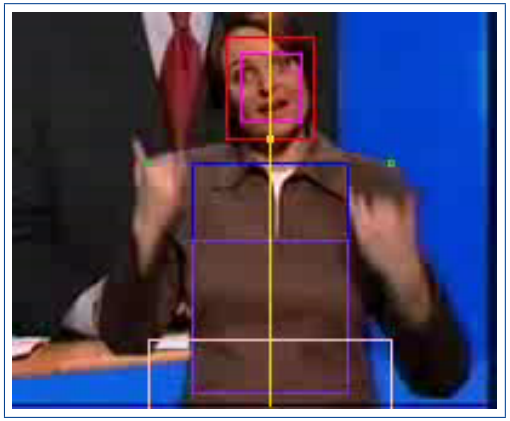
System Overview - Tracking

- **Examples**



overlay of two consecutive frames with labeled hand positions

head and hand tracking on RWTH-Boston-104 database



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

- **further information about tracking framework and scoring functions: in FGR 2006 paper**

System Overview - Composite Features

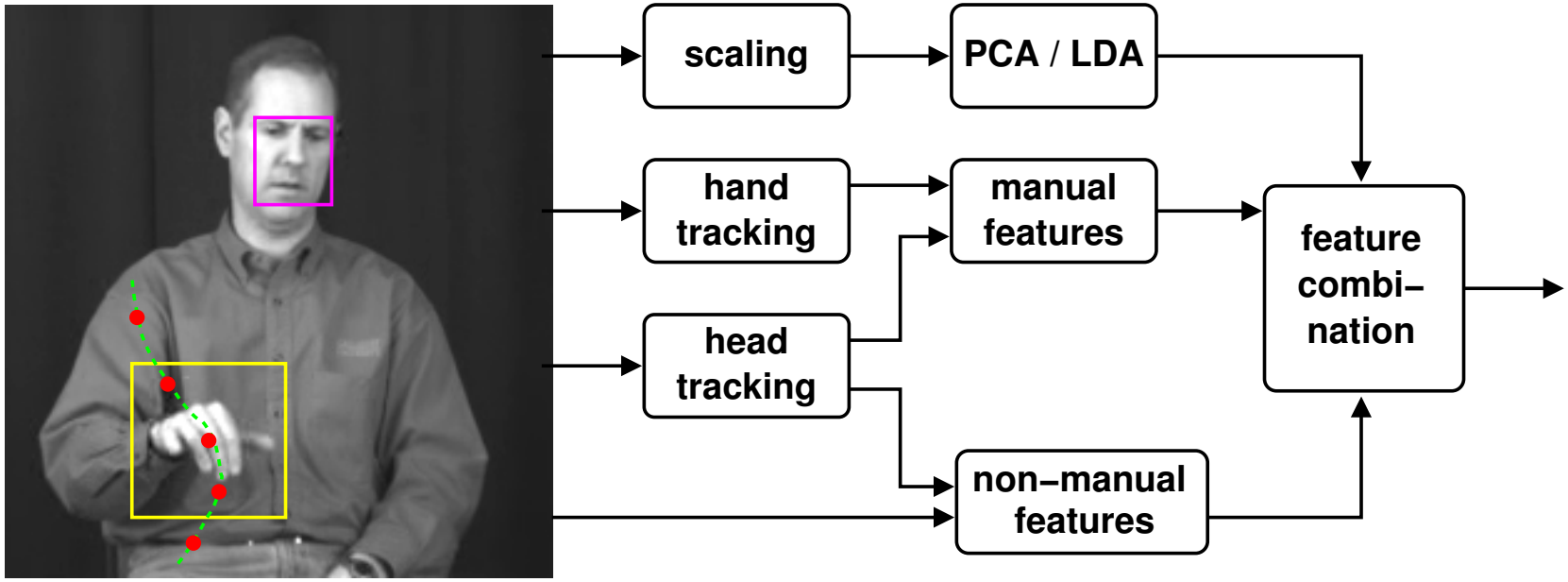
- Observation: border-pixels have low or zero variance
- Goal: reduce dimensionality of input data (32 $\times$ 32 pixel frames)

- LDA : not enough data to estimate scatter matrices correctly
- PCA : more stable for high dimensional data and small training sets
- WINDOWING :
  - often applied in speech recognition
  - feature vectors in a sliding window are combined to one feature vector $[\ldots, x_{t-1}, x_t, x_{t+1}, \ldots]^T$, which is then transformed with a LDA to find an optimal combination of these features
  - for an increasing window size an increasing amount of training data is needed

$\Rightarrow$PCA + [WINDOWING] + [LDA | PCA]

System Overview - Composite Features

Composite Features using FLOW network



- **Manual features (using tracking):**
  - **Hand position u_t, and motion $m_t = u_t - u_{t-\Delta}$, $\Delta \in \{1, 2, \dots\}$**
  - **Hand trajectory**
- **Non-manual features:**
  - **Motion $e_t = \sqrt{\sum_{u \in U} \left(\boldsymbol{X}_t[u] - \boldsymbol{X}_{t-1}[u]\right)^2}$**

System Overview

Further consolidated findings from speech and image recognition

- **Context dependent phonemes: Phoneme in diphone / triphone context**
- **Across-word models: Triphone context at word boundaries**

- **Model Image Variablility using Tangent Distance or IDM to allow global or local transformations**
- **Prononciation handling using a modified lexicon and prononciation weighting**
- **Speaker adaptive system using e.g. MLLR techniques**
- **Recognition**
  - **feature combination vs. model combination**
  - **discriminative model combination**

⇒available in SPRINT

Databases & Experimental Results

RWTH-Boston-104 Database (RWTH Aachen University)

- Data published by National Center for Sign Language and Gesture Resources of the Boston University
- Collected mainly for research on syntactic structure of American sign language
- 201 annotated American sign language sentences
- Vocabulary: 104 words, 3 speakers (2 ♀, 1 ♂)

JOHN WRITE HOMEWORK

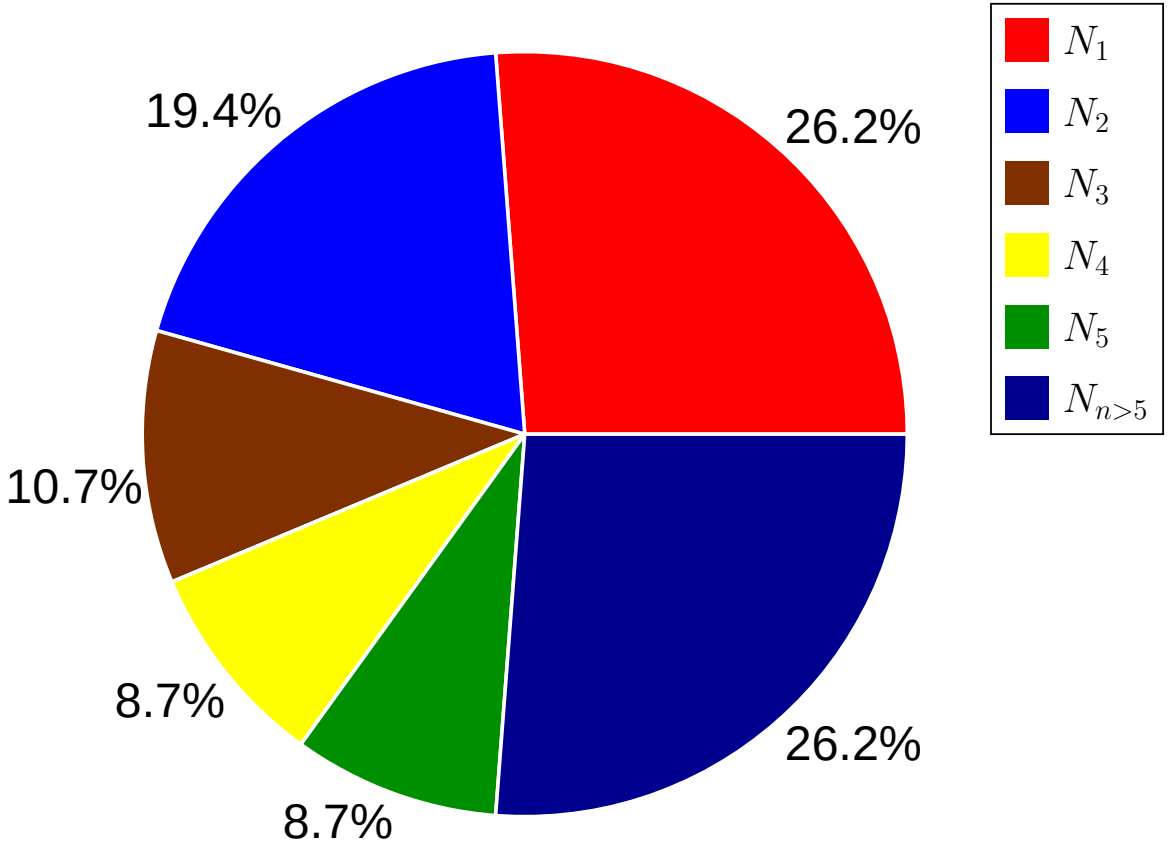
LIKE CHOCOLATE WHO

SOMETHING/-ONE CAR STOLEN

Databases & Experimental Results

RWTH-Boston-104: Statistics

corpus	sentences	running glosses	unique glosses	singletons	OOV signs
training	161	710	103	27	-
test	40	178	65	9	1

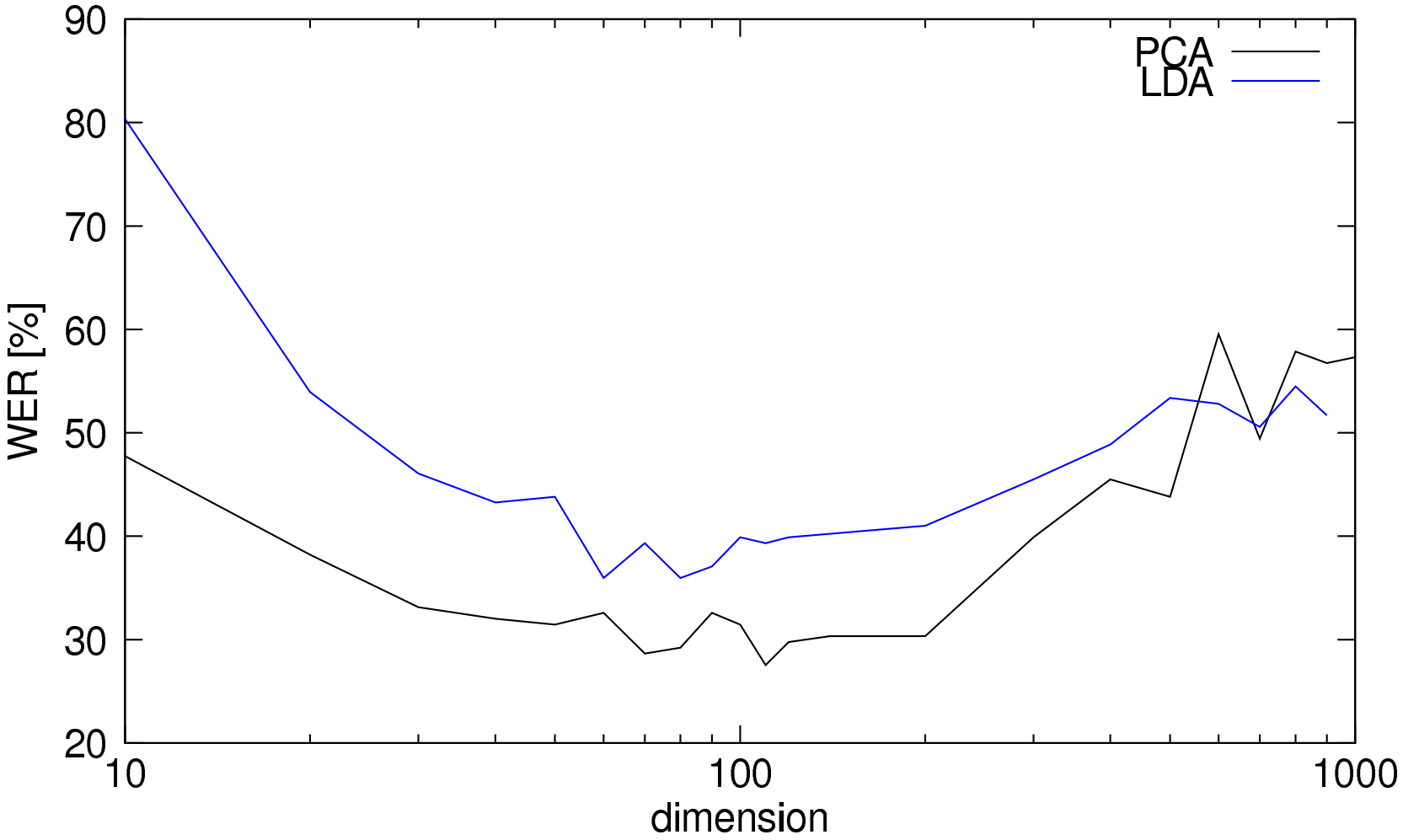


word counts in training corpus. $N_n := |\{w : N_w = n\}|$

LM type	perplexity PP
zerogram	106.0
unigram	36.8
bigram	6.7
trigram	4.7

Databases & Experimental Results

RWTH-Boston-104: Results Linear Mappings



results for LDA / PCA transformed frames

Databases & Experimental Results

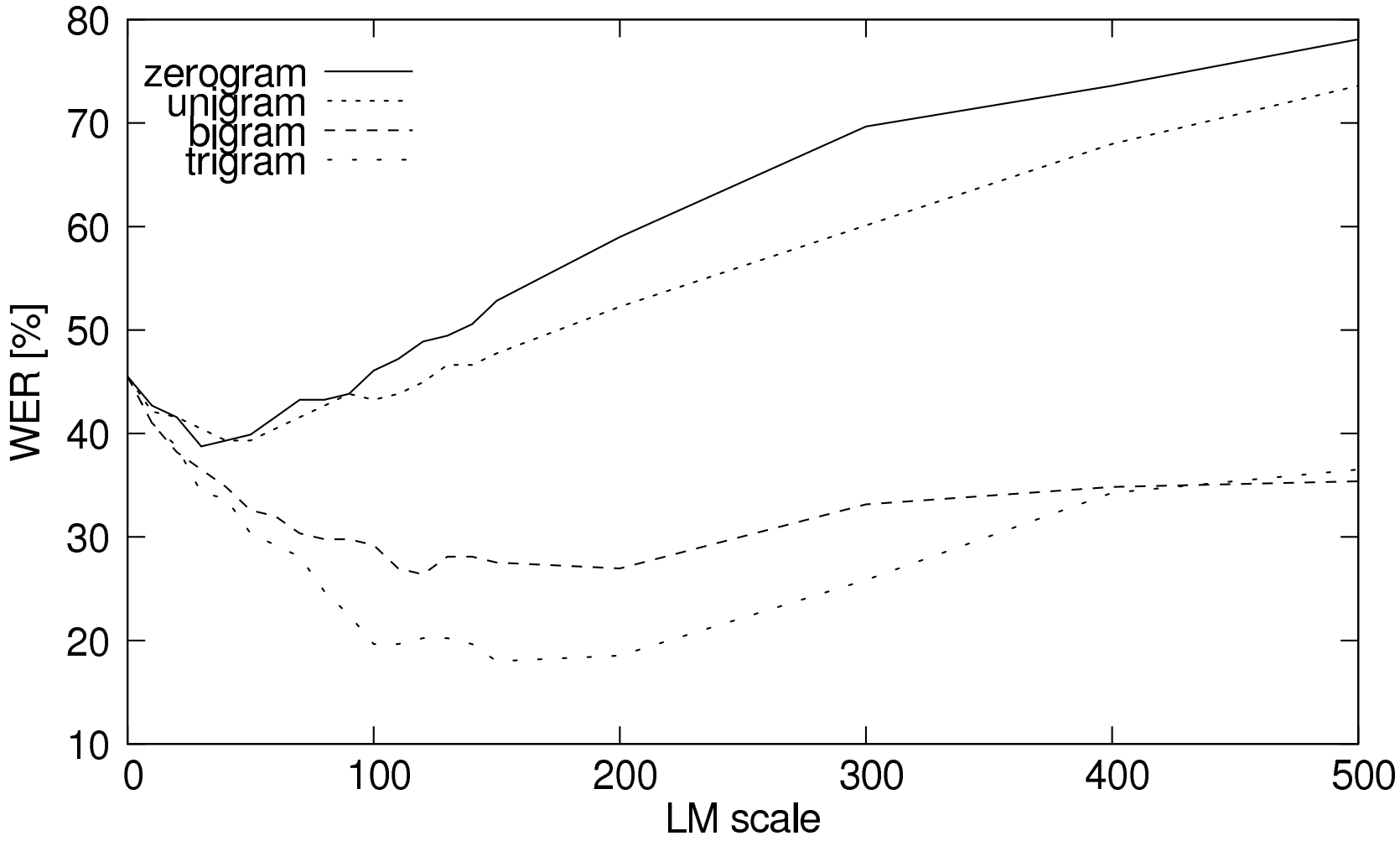
RWTH-Boston-104: Results Linear Combination

- **Setup:**
  - **32×32 frames PCA transformed to 110 components (PCA-110)**
  - **LDA transformation to 80 components**

window size	WER[%]
1	30.3
3	26.4
5	**25.8**
7	30.3

Databases & Experimental Results

RWTH-Boston-104: Language Modeling



results for different LM and scales

Databases & Experimental Results

RWTH-Boston-104 Database: Results feature & model combination

Word Error Rates in [%] on the RWTH-Boston-104 database

Feature Combination	WER[%]
intensity 32x32	**37.1**
PCA-110	**21.9**
PCA-110+Windowing-5+PCA-100	**20.8**
hand position	**59.6**
hand motion	**56.7**
hand position+motion	**48.9**
hand trajectory	**73.6**
PCA-110 +hand trajectory	**20.2**
Model Combination	**WER[%]**
PCA-110 + trajectory & PCA-110 + motion	**17.9**

Databases & Experimental Results

RWTH German Fingerspelling Database (RWTH Aachen University)

- **Color images (320x240 and 352x288), non-uniform lighting, no user constraints**
- **20 persons, 2 sessions (700 train and 700 test sequences)**
- **Classification task: 35 dynamic gestures**

Error Rates in [%]

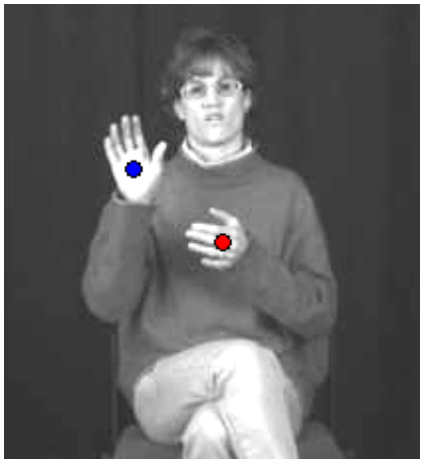
Features	Euclidian	Tangent
intensity, skin color thresholding	87.1	-
+ camshift tracker (no segmentation)	44.0	35.7

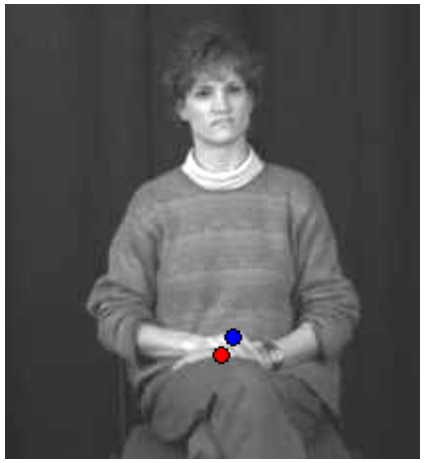
⇒**available at** `http://www-i6.informatik.rwth-aachen.de/~dreuw/`

Databases & Experimental Results

RWTH-Boston-Hands Database (RWTH Aachen University)

- Currently 15 videos (1119 frames)
- Self annotated positions of right and left hand





Databases & Experimental Results

RWTH-Phoenix Database (RWTH Aachen University)

- Weather forecast in ARD Tagesschau recorded from Phoenix
- Currently: 95 videos recorded between 2004-04-08 and 2005-12-21
- Fully annotated GSL with ELAN, 1353 sentences, 9455 running words
- Vocabulary: 711 words, 11 speakers (10 ♀, 1 ♂)
- Complex, changing background
- Virtual signing space (references & verb flexion)


- see also LREC 2006
- http://www.mpi.nl/tools/elan.html

Outlook

- Experiments on larger databases:
  - RWTH-Boston-447: 447 words, 890 sentences, 4805 running words, 5 speakers (2 ♂, 3 ♀), $PP \approx 30$
  - RWTH-Phoenix
- Combination of composite feature vectors
  - Feature Combination
  - Discriminative Model Combination DMC
  - ROVER using confidence measures of different output alignments
- Features
  - integrate probabilistic body model
  - facial feature analysis
  - virtual signing space analysis
- create new and publicly available benchmark corpora??
- find university partners??
- build a robust sign language recognition system

Thank you for your attention

Philippe Dreuw

dreuw@cs.rwth-aachen.de

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

References

[Bungeroth & Stein+ 06] J. Bungeroth, D. Stein, P. Dreuw, M. Zahedi, H. Ney: A German Sign Language Corpus of the Domain Weather Report. In *Fifth International Conference on Language Resources and Evaluation*, accepted, Genoa, Italy, May 2006. 29

[Dreuw & Deselaers+ 06a] P. Dreuw, T. Deselaers, D. Keysers, , H. Ney: Modeling Image Variability in Appearance-Based Gesture Recognition. In *3rd Workshop on Statistical Methods in Multi-Image and Video Processing (ECCV-SMVP)*, accepted, Graz, Austria, May 2006. 29

[Dreuw & Deselaers+ 06b] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, H. Ney: Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition. In *7th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2006*, IEEE, pp. 293–298, Southampton, April 2006. 29

[Haeb-Umbach & Ney 92] R. Haeb-Umbach, H. Ney: Linear Discriminant Analysis for Improved Large Vacabulary Continuous Speech Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 13–16, 1992. 29

[Kanthak & Molau+ 00] S. Kanthak, S. Molau, A. Sixtus, R. Schlüter, H.Ney: The RWTH Large Vocabulary Speech Recognition System for Spontaneous Speech. In *Proceedings of the Konvens 2000*, pp. 249–254, Ilmenau, Germany, 2000. 29, 30

[Katz & Meier+ 02] M. Katz, H.G. Meier, H. Dolfing, D. Klakow: Robustness of Linear Discriminant Analysis in Automatic Speech Recognition. In *International Conference on Pattern Recognition*, Vol. 3, pp. 371–374, 2002. 29

[Martinez & Kak 01] A. Martinez, A. Kak: PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 228–233, Feb. 2001. 29

[Neidle & Lee 00] J.K.D.M.B.B. Neidle, C., R. Lee: *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. MIT Press, Cambridge, MA, 2000. 29

[Woodland 01] P.C. Woodland: Speaker Adaptation for Continuous Density HMMs: A Review. In *Adaptation Methods for Speech Recognition*, pp. 11–19, Sophia Antipolis, France, Aug. 2001. 29

[Zahedi & Keysers+ 05a] M. Zahedi, D. Keysers, T. Deselaers, H. Ney: Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition. In *DAGM 2005, Pattern Recognition, 26th DAGM Symposium*, number 3663 in Lecture Notes in Computer Science, pp. 401–408, Vienna, Austria, August 2005. 29

[Zahedi & Keysers+ 05b] M. Zahedi, D. Keysers, T. Deselaers, H. Ney: Pronunciation Clustering and Modeling of Variability for Appearance-Based Sign Language Recognition. In *6th International Workshop on Gesture in Human-Computer Interaction and Simulation*, number 3881 in Lecture Notes in Artificial Intelligence, pp. 68–79, Vienna, Austria, May 2005. 29

Appendix: Literature

- **SPRINT : [Kanthak & Molau+ 00]**
- **Tracking: [Dreuw & Deselaers+ 06b]**
- **LDA/PCA:**
  - **[Haeb-Umbach & Ney 92]**
  - **[Martinez & Kak 01]**
  - **[Katz & Meier+ 02]**
- **Tangent/IDM: [Dreuw & Deselaers+ 06a], [Zahedi & Keysers+ 05a]**
- **Prononciation: [Zahedi & Keysers+ 05b]**
- **Speaker Adaption: [Woodland 01]**
- **Boston data: [Neidle & Lee 00]**
- **Sign language corpora: [Bungeroth & Stein+ 06]**

Appendix: Infrastructure

- Software: SPRINT [Kanthak & Molau+ 00] written in C++, highly optimized
- Computing power
  - local: 100 nodes (including desktops), +130 computing nodes
  - 20TB raid storage system
  - external: RWTH high performance computing center (several SunFire and AMD-Opteron clusters)
- Queueing system / clustermanagement:
  - SUN Grid Engine
  - batch mode processing of computing requests
  - parallel processing possible
  - all experiments can be submitted to local or external queueing system
  - allows to distribute training/recognition onto multiple nodes

Appendix: Modeling

Word modelling

- Words as a sequence of HMM-States

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

- Super HMMs for sentence w_1^N

$$\Pr(x_1^T, s_1^T|w_1^N) = \prod_{t=1}^{T} p(x_t, s_t|s_{t-1}, w_1^N)$$

- Word models

$$p(x_t, s|s', w) = p(s|s', w) \cdot p(x_t|s, w)$$

Language model

- Trigram:

$$\Pr(w_1^N) = \prod_{n=1}^{N} p(w_n|w_{n-2}, w_{n-1})$$

Appendix: Eigenfaces

- An image X can be projected to face space by a linear transformation ϕ:

$$\phi(X) = V^T(X - \mu)$$

  where $V = [v_1 \dots v_m]$ is the matrix of the first m eigenvectors and μ is the mean face calculated on the set of training images.

- The projection from face space to image space is:

$$\phi^{-1}(X_f) = VX_f + \mu$$

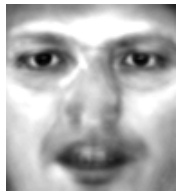
  where X_f is the image representation in face space $\phi(X)$.

- The distance between an image and its forward and backward projected version, is called the *face space distance*. It can be used as a measure of "faceness".

$$d_f(X) = \|X - \phi^{-1}(\phi(X))\|^2$$

Appendix: Eigenfaces

- **An example of projected images and the resulting distance:**

X	$\phi^{-1}(\phi(X))$	$X - \phi^{-1}(\phi(X))$	$d_f(X)$
			278
			432

- **We use the face space distance as score function to detect and track heads:**

$$s_f(u_{t-1}, u_t; X_{t-1}^t) = -d_f\left(X_t(u_t)\right)$$

where $X_t(u_t)$ denotes a rectangular patch of image X_t centered in position u_t.

Appendix: LM Perplexity

- The perplexity of a language model and a test corpus w_1^N is defined as:

$$\begin{aligned} PP &= p(w_1^N)^{-\frac{1}{N}} \\ &= \left[\prod_{n=1}^{N} p(w_n|h_n)\right]^{-\frac{1}{N}} \end{aligned}$$

- As the perplexity is an inverse probability, it can be interpreted as the average number of possible words at each position in the text.
- The logarithm of the perplexity is equal to the entropy of the text, i.e. the redundancy of words in the test corpus with respect to this language model.

$$\log PP = -\frac{1}{N} \sum_{n=1}^{N} \log p(w_n|h_n)$$

Appendix: LM Scales

- In bayes' decision rule, acoustic model and language model have the same impact on the decision.
- Experiments in speech recognition have shown that the recognition performance can be improved, if the language model has more weight than the acoustic model.
- The weighting is done by introducing a language model scale α and an acoustic model scale β:

$$\begin{aligned}\underset{w_1^N}{\operatorname{argmax}} \left\{ p(w_1^N|x_1^T) \right\} &= \underset{w_1^N}{\operatorname{argmax}} \left\{ p^\alpha(w_1^N) \cdot p^\beta(x_1^T|w_1^N) \right\} \\ &= \underset{w_1^N}{\operatorname{argmax}} \left\{ \frac{\alpha}{\beta} \log p(w_1^N) + \log p(x_1^T|w_1^N) \right\}\end{aligned}$$

- The factor $\frac{\alpha}{\beta}$ is referred to as *language model factor*.

Appendix: Hand Trajectory Features

- calculate global features describing geometric properties of the hand trajectory
- estimation of the covariance matrix Σ_t for hand positions in a certain time window $2\Delta + 1$

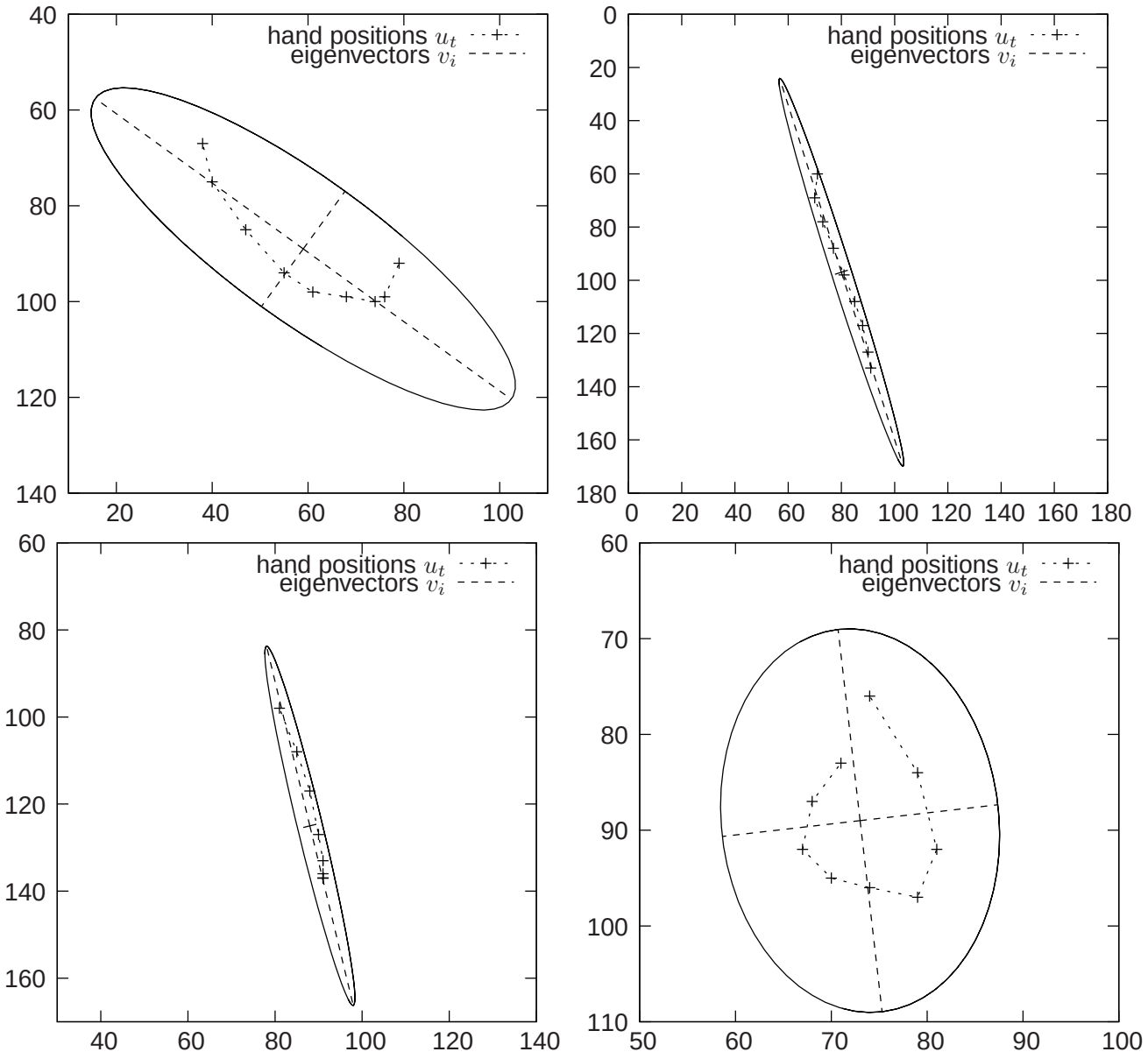
$$\mu_t = \frac{1}{2\Delta + 1} \sum_{t'=t-\Delta}^{t+\Delta} u_{t'}$$

$$\Sigma_t = \frac{1}{2\Delta + 1} \sum_{t'=t-\Delta}^{t+\Delta} (u_{t'} - \mu_t) (u_{t'} - \mu_t)^T$$

$$\Sigma_t v_{t,i} = \lambda_{t,i} \cdot v_{t,i} \qquad i \in \{1, 2\}$$

- eigenvalues $\lambda_{t,i}$ and eigenvectors $v_{t,i}$ of the covariance matrix can then be used as global features.

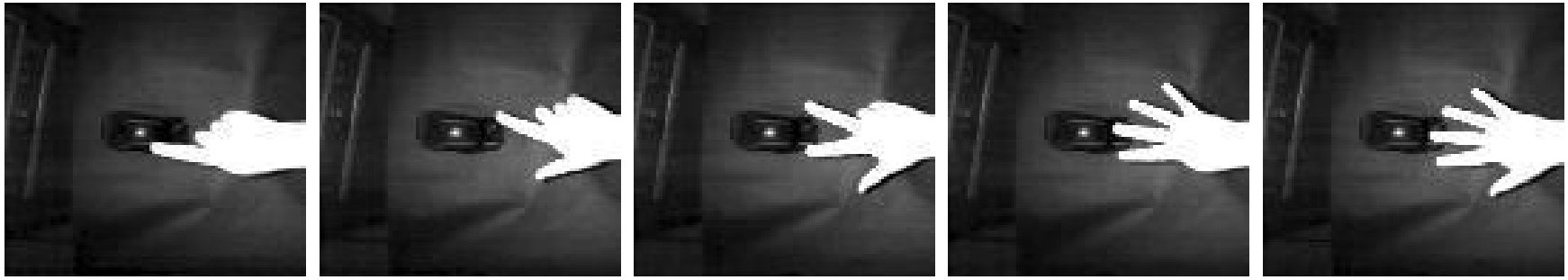
Appendix: Hand Trajectory Features



Appendix: Gesture Recognition

LTI-Gesture Database (RWTH Aachen University)

- **Infrared gray images (106x96), top view of signing hand**
- **140 train- and 140 test sequences**
- **Classification task: 14 dynamic gestures**



best error rate: 0% ER

Appendix: Gesture Recognition

DUISBURG-Gesture Database (Duisburg University)

- Gray images (96x72 pixels), full body-centered space
- 14 persons, 336 sequences
- Classification task: 24 dynamic gestures, leaving-one-person-out

best error rate: 10% ER