Kontinuierliche Gebärdenspracherkennung auf großem Vokabular

Philippe Dreuw, Morteza Zahedi, David Rybach, Thomas Deselaers, Hermann Ney

dreuw@informatik.rwth-aachen.de

Gebärdensprachworkshop 27. Oktober 2006

Lehrstuhl für Informatik 6
RWTH Aachen University, Germany

Übersicht

1 Einleitung

2 Systemübersicht

3 Wortmodellierung

4 Ergebnisse

5 Schlussfolgerung

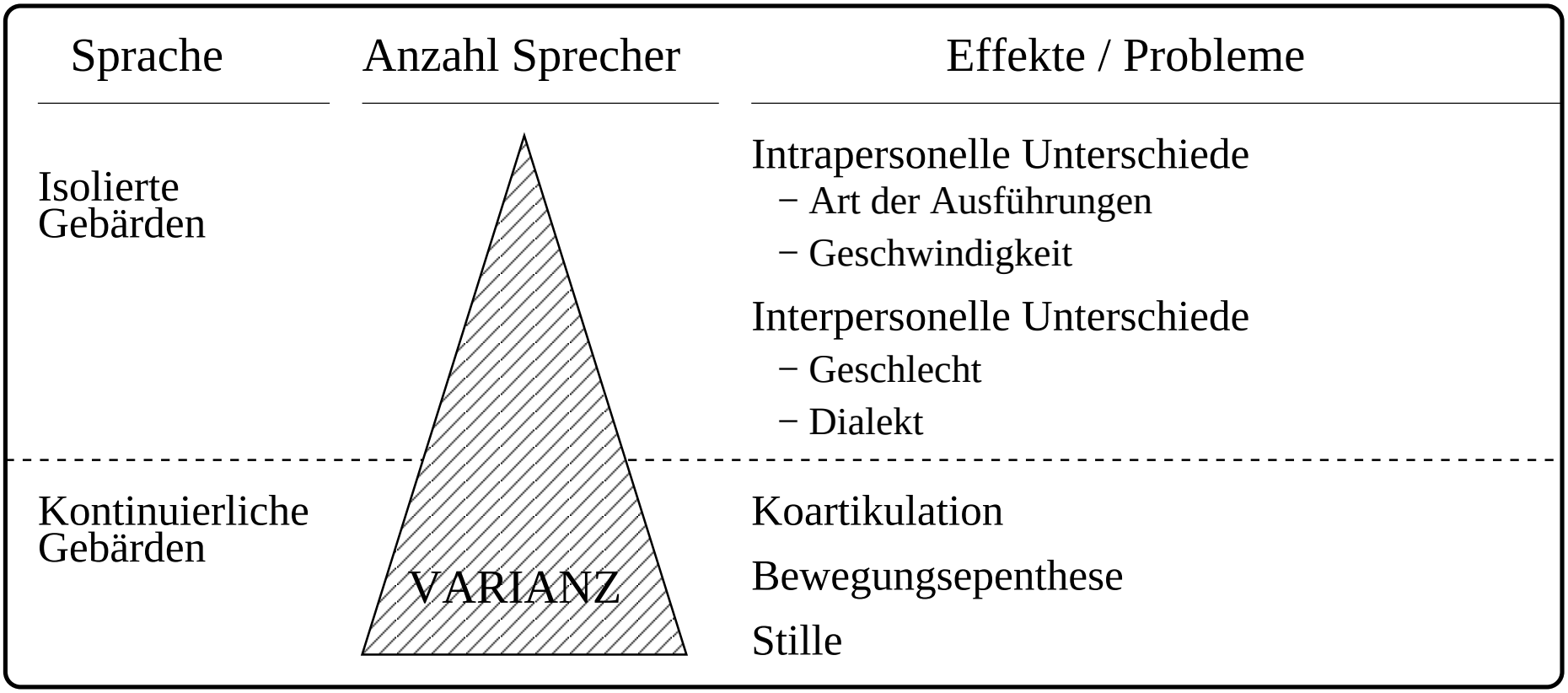
1 Einleitung

- **Gemeinsamkeiten: Spracherkennung / Gebärdenspracherkennung**

Lautsprache	$\Longleftrightarrow$	zeitliche Folge von Lauten
Gebärdensprache		**zeitliche Folge von Gebärden / Gesten**
$\Downarrow$		$\Downarrow$
akustisches Signal	$\Longleftrightarrow$	zeitliche Folge von akustischen Vektoren, (akustische Realisierung der Laute)
visuelles Signal		**zeitliche Folge von "visuellen" Vektoren, (visuelle Realisierung der Gebärden)**

- **Unterschiede:**
  - **Grammatik**
  - **Parallelität der Gebärdensprache**
  - **Raumnutzung und Indexierung**

Problematik robuster Erkennungssysteme



- **Was ist in der Spracherkennung wesentlich anders? Was fehlt uns?**
  - **Daten: ca. 400 Stunden Sprache vs. $<$ 1 Stunde Gebärdensprache**
  - **Sprecher: 50 – 100 Sprecher vs. 1 – 4 Gebärdensprecher**
  - **robuste Merkmale, Phoneme, Modelle, ...**

Spracherkennung: Geschichte

- **Wie lange hat Spracherkennung gedauert?**

zeitliche Entwicklung (Forschung):

1965	erste Versuche
1975	Einzelwörter: isoliert gesprochen, kleiner Wortschatz
1985	isoliert: 5000 Wörter kontinuierlich: 1000 Wörter
1990	kontinuierlich: 10 000 Wörter
1995	Telefonsysteme: sprecherunabhängig, kontinuierlich, 3 000 Wörter

Gebärdenspracherkennung: Geschichte

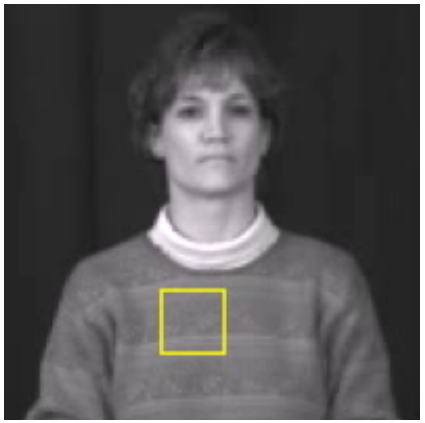
- **Wo stehen wir heute im Vergleich zur Spracherkennung?**

zeitliche Entwicklung (Forschung):

1965	Beschreibung möglicher Wortuntereinheiten
1975	Beschreibung nicht-manueller Komponenten
1990	Beschreibung der Bewegungsepenthese, HamNoSys
1995	isoliert, Kamera: 40 Wörter
2000+	isoliert, Handschuhe: 10 - 5000 Wörter
	kontinuierlich, Kamera: 10 - 100 Wörter

Anwendung: Sprache-zu-Sprache

Erkennung: Sprache-zu-Text (Video $\Rightarrow$Glossen)



$\Downarrow$

Übersetzung: Text-zu-Text (Glossen $\Rightarrow$Text)

`JOHN FISH WONT EAT BUT CAN EAT CHICKEN`

John will not eat fish but eats chicken

$\Downarrow$

Synthese: Text-zu-Sprache (Text $\Rightarrow$Audio)

`audio/021.wav`

2 Systemübersicht

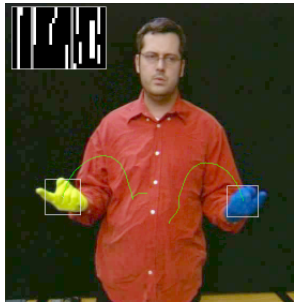
- **Erkennung**
- Machine Translation
- Synthese

2.1 Merkmale zur Gebärdenspracherkennung

- **Was für Merkmale brauchen wir?**
  - **Manuelle Komponenten**
  - **Nicht-Manuelle Komponenten**

  ⇒ sollten in irgendeiner Form aus dem Eingangssignal extrahiert werden

- **Unterschiedliche Ansätze / Annahmen**
  - **Spezial Hardware**
  - **Computersehen (Computer Vision)**


⇒ dabei Entstehen unterschiedliche Probleme bei der Extraktion

2.2 Datenbank

RWTH-Boston-104 (RWTH Aachen University)

- American Sign Language (ASL)
- Daten der Boston University
- 201 annotierte ASL Sätze
- Vokabular: 104 Wörter, 3 Sprecher (2 ♀, 1 ♂)

JOHN WRITE HOMEWORK

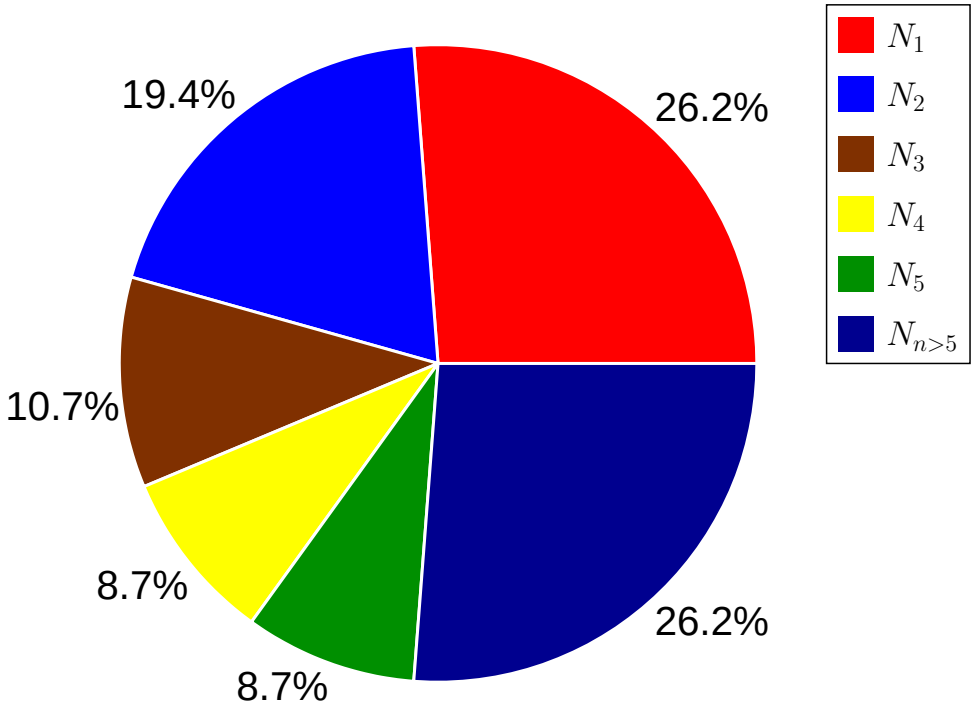
LIKE CHOCOLATE WHO

SOMETHING/−ONE CAR STOLEN

Datenbank

RWTH-Boston-104: Statistik

Korpus	Sätze	Glossen	Vokabular	Einzelbeobachtungen
Training	161	710	103	27
Test	40	178	65	9



Worthäufigkeiten im Training

2.3 Tracking

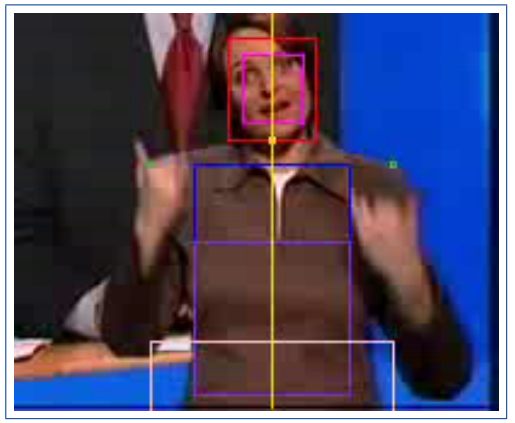
- **Anwendung: Tracking des Kopfes**
- **Probleme:**
  - **Hände sind oft vor dem Gesicht**
  - **Kopf Rotation, starke Mimik**
  - **Hintergrund: Hautfarben, Struktur, ...**

- **Idee: kombiniere Hautfarbinformation und Gesichtsmerkmale**

Tracking

- **Beispiele**

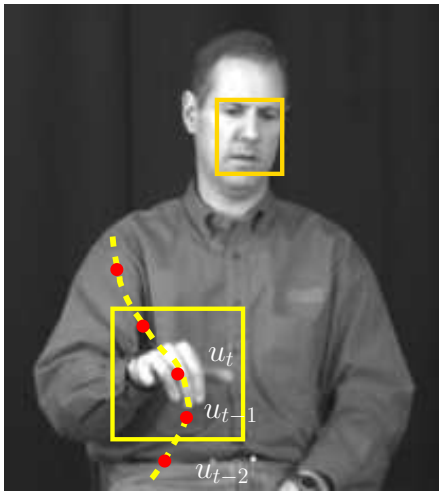
Kopf- und Hand-Tracking auf der RWTH-Boston-104 Datenbank



Kopf-Tracking auf der RWTH-Phoenix Datenbank mit Körpermodell

2.4 Verwendete Merkmale

- **Manuelle Merkmale (aus dem Tracking):**
  - **Hand-Position**
  - **Hand-Bewegung**
  - **Hand-Trajektorie**

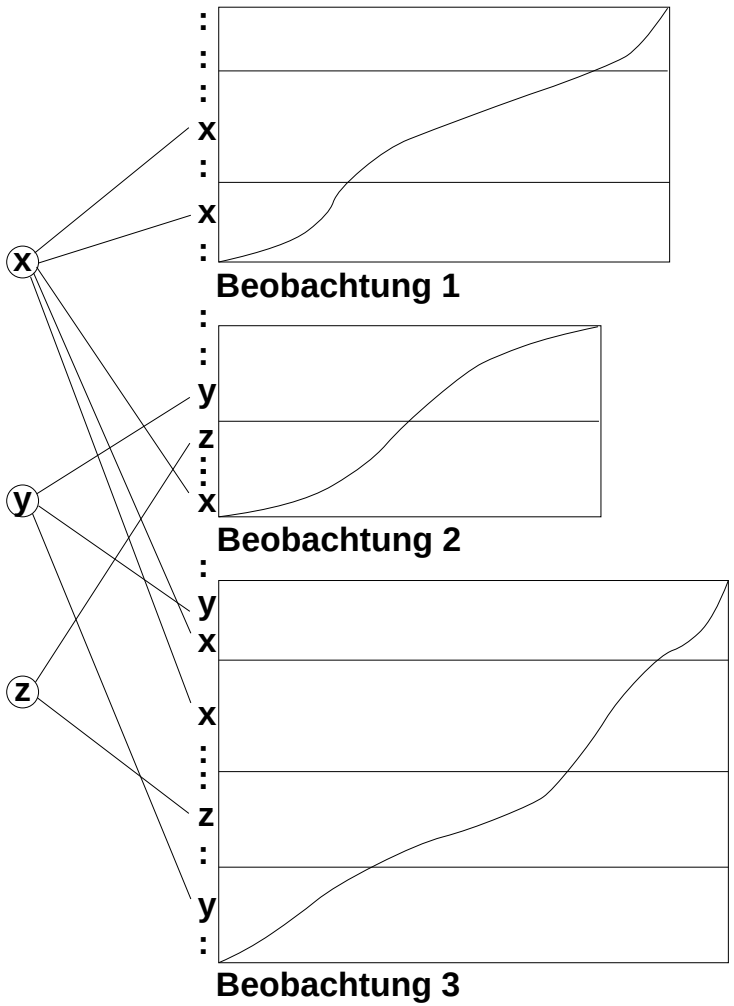


Baseline Setup:

- **Erscheinungsbasierte Bildmerkmale**
  - **Bilder werden auf 32×32 Pixel verkleinert**
  - **dienen als gutes Baseline Ergebnis in zahlreichen Problemen der Bilderkennung**
  - **erfolgreich in der Gestenerkennung eingesetzt**

3 Wortmodellierung

- **Erkennung auf großem Vokabular: Ganzwort-Modelle nicht sinnvoll**
  - **nicht ausreichend Trainingsmaterial vorhanden**
  - **größerer Speicheraufwand**
- **Lösung:**
  - **erstelle Wort-Modelle durch Konkatenierung von Wortuntereinheiten**
- **Vorteile:**
  - **Daten werden unter den Wörtern aufgeteilt**
  - **dadurch mehr Trainingsdaten pro Wort**
  - **nicht im Training gesehene Wörter können nun durch ein Aussprache-Lexikon erkannt werden**



Wortmodellierung

- **Probleme in der Gebärdenspracherkennung:**
  - **Phoneme noch immer nicht eindeutig definiert**
  - **kein Aussprache-Lexikon vorhanden**
  - **Phoneme treten simultan auf (Multi-Stream)**
  - **deutlich mehr Phoneme in der Gebärdenspracherkennung als Phoneme in der Spracherkennung**

$\Rightarrow$ **Ansatz nicht ohne weiteres auf die Gebärdenspracherkennung übertragbar**

- **Isolierte Gebärden**
  - **Wortgrenzen bekannt**

- **Kontinuierliche Gebärden**
  - **unbekannte Wortgrenzen**
  - **Kontexteffekte an Wortübergängen**
  - **Bewegungsepenthese**
  - **Stille**

4 Ergebnisse

- **Baseline Ergebnisse und Kombination mit Hand Merkmalen**

Merkmal	Fehlerrate
skaliertes Bild	37.0
PCA-transformiertes Bild	27.5
+Hand-Trajektorie	23.6
Fenstern	21.9
Modell-Kombination	17.9

- **Beispielsätze**

ALL	BOY	GIVE	TEACHER	APPLE
ALL	BOY	GIVE	TEACHER	APPLE
JOHN	SHOULD	NOT	BUY	HOUSE
JOHN	FUTURE	NOT	BUY	HOUSE
ANN	BLAME	MARY		
ANN	BLAME	_____		
JOHN			READ	BOOK
JOHN	FUTURE	FINISH	READ	BOOK

5 Schlussfolgerung

- Ergebnisse wurden auf einer öffentlichen Datenbank erzielt
- System benötigt keine spezielle Hardware oder Handschuhe
- Gebärdenspracherkennung mit einem aktuellen Spracherkennungssystem
- erscheinungsbasierte Bildmerkmale erzielen auch gute Ergebnisse in der Gebärdenspracherkennung
- viele Prinzipien der Spracherkennung sind direkt auf die Gebärdenspracherkennung übertragbar
  - besonders wichtig: Kontextinformation und Sprachmodelle

Ausblick

- **Weitere Erkenntnisse der Sprach- und Bilderkennung untersuchen:**
  - **Sprecheradaption**
  - **weitere Merkmale für die Erkennung**
  - **Wortmodellierung**

- **Integration der Rauminformation aus der Erkennung in die Übersetzung**

Erkennung	**JOHN IX GIVE MAN IX NEW COAT** **JOHN __ GIVE _____ IX NEW COAT**
Übersetzung ohne Rauminformation	**John gives that man a coat**
Übersetzung mit Rauminformation	**John gives the man over there a coat.**

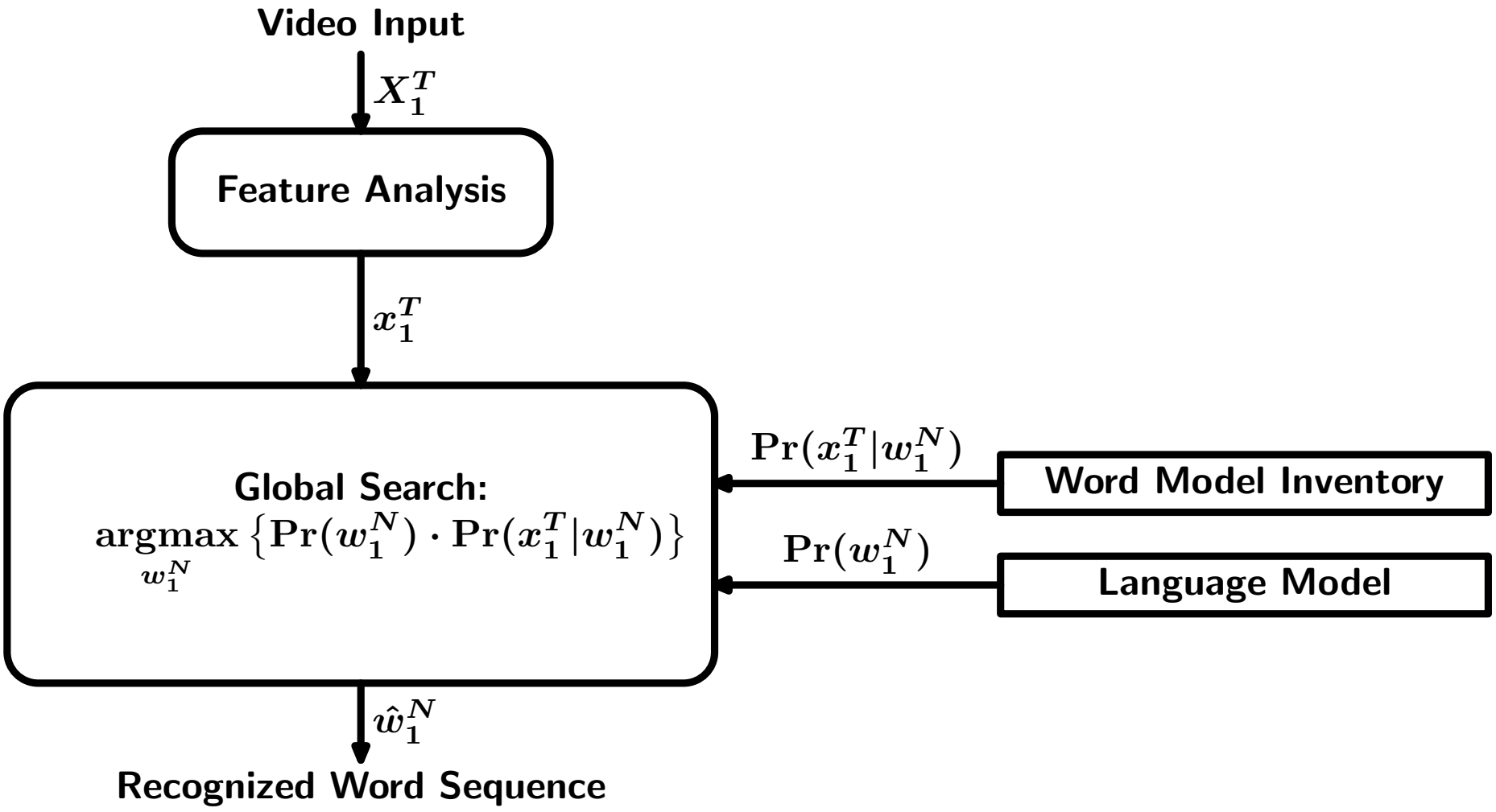
Danke für Ihre Aufmerksamkeit

Philippe Dreuw

dreuw@informatik.rwth-aachen.de

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

Anhang: Bayes'sche Entscheidungsregel



Anhang: 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:

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

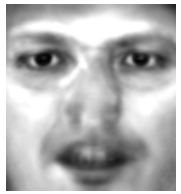
  where X_f is the image representation in face space $\phi(X)$.

- 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".

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

Anhang: 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.

Anhang: LM Scales

- **Akustisches Modell und Sprachmodell haben den gleichen Einfluss in der Bayes'schen Entscheidungsregel**
- **Experimente in der Spracherkennung haben gezeigt, dass die Erkennung stark verbessert werden kann, wenn das Sprachmodell einen stärkeren Einfluss als das Akustische Modell hat**
- **Die Gewichtung erfolgt durch die Einfühlung eines Gewichtes α für das Sprachmodell und eines Gewichtes β für das Akustische Modell:**

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

- **Der Faktor $\frac{\alpha}{\beta}$ wird als Sprachmodell Faktor bezeichnet.**

Anhang: 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.

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

Anhang: 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 $2\Delta + 1$

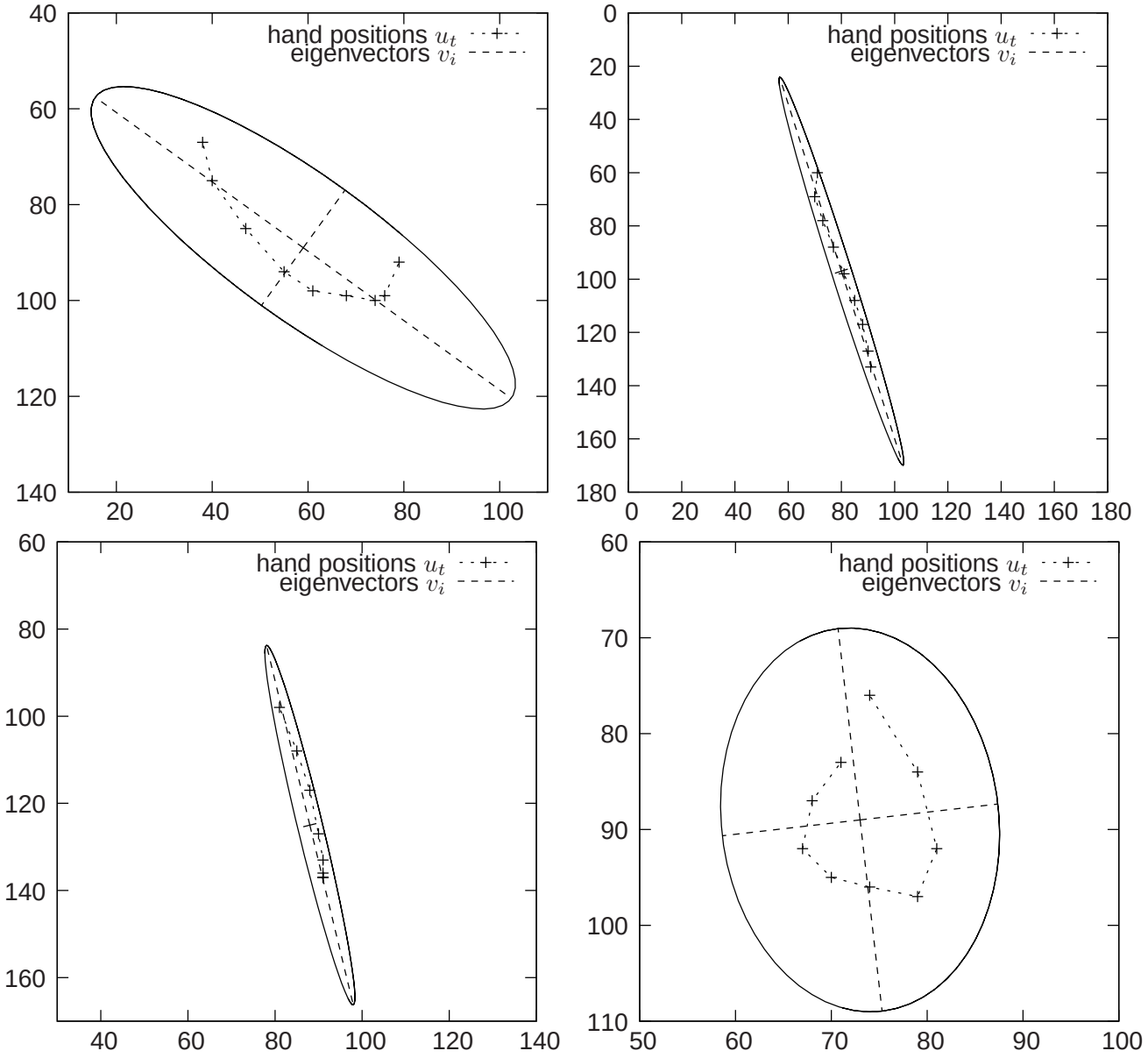
$$\mu_t = \frac{1}{2\Delta + 1} \sum_{t'=t-\Delta}^{t+\Delta} u_{t'}$$

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

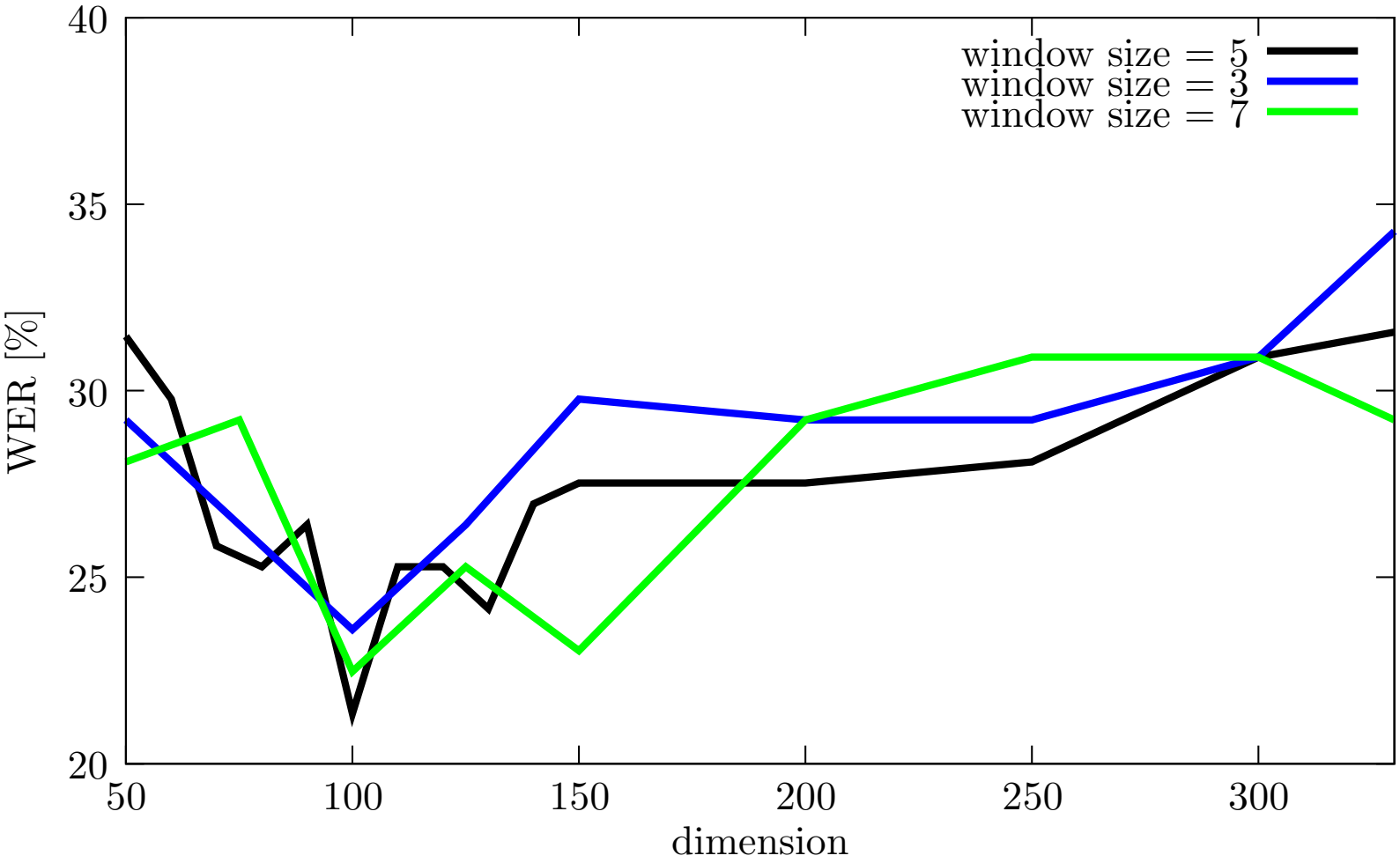
$$\Sigma_t v_{t,i} = \lambda_{t,i} \cdot v_{t,i} \qquad i \in \{1, 2\}$$

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

Anhang: Hand Trajectory Features

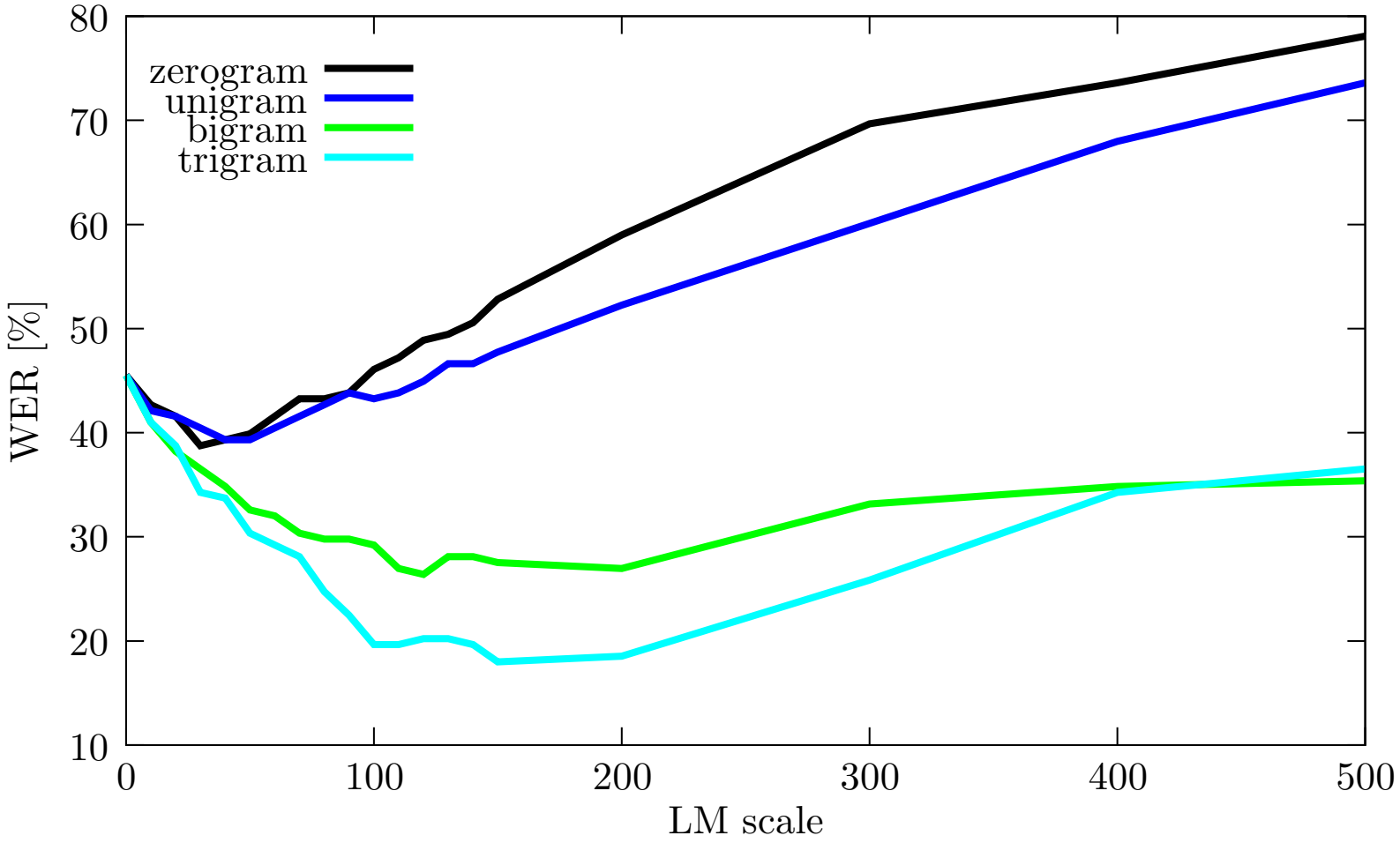


Anhang: Ergebnisse - Kontextinformation



Kombination durch Fenstern von PCA-transformierten Bildern

Anhang: Ergebnisse - Sprachmodellierung



Ergebnisse für unterschiedliche Sprachmodelle und Skalierungsfaktoren