Probabilistic Sequence Models for Image Sequence Processing and Recognition

Final PhD Talk

Philippe Dreuw

Lehrstuhl für Informatik 6
Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University
D-52056 Aachen, Germany

Apr. 27th, 2012

Outline

Introduction

RWTH AACHEN UNIVERSITY

Introduction

What is the Problem?

Why is it Difficult?

How can we handle all these Challenges?

Introduction

What is the Problem?

- we want to recognize continuous symbol streams
  - character, syllable, word, gesture, sign, ...

Why is it Difficult?

How can we handle all these Challenges?

Introduction

What is the Problem?

- we want to recognize continuous symbol streams
  - character, syllable, word, gesture, sign, ...

Why is it Difficult?

- high variability of the signal
  - appearance typically varies over time
  - realized within an important spatio-temporal context
- most decisions are interdependent
  - symbol boundaries are not always visible
- natural samples
  - inter- and intra-personal variability
  - hesitations, dialects, styles, genres, ...

How can we handle all these Challenges?

Introduction

What is the Problem?

- we want to recognize continuous symbol streams
  - character, syllable, word, gesture, sign, ...

Why is it Difficult?

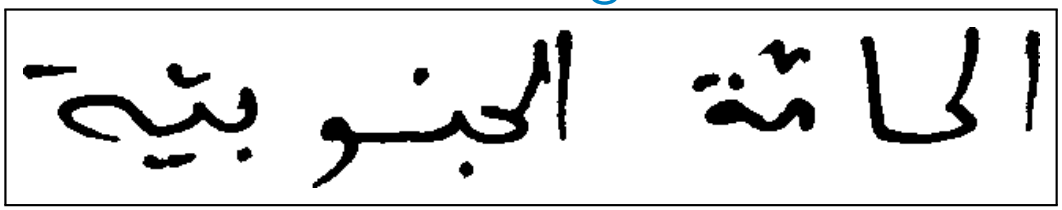
- high variability of the signal
  - appearance typically varies over time
  - realized within an important spatio-temporal context
- most decisions are interdependent
  - symbol boundaries are not always visible
- natural samples
  - inter- and intra-personal variability
  - hesitations, dialects, styles, genres, ...

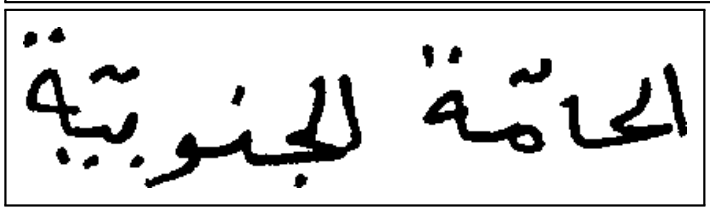
How can we handle all these Challenges?

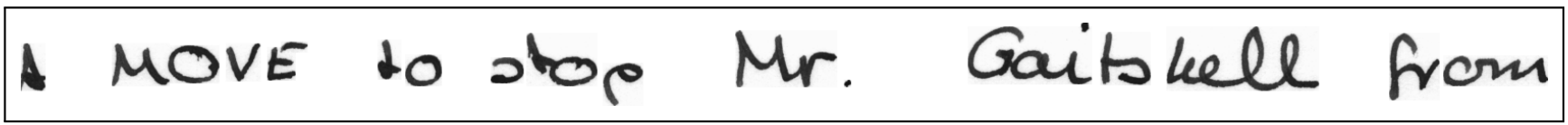
$\Rightarrow$ HMM based approaches are probably the method of choice

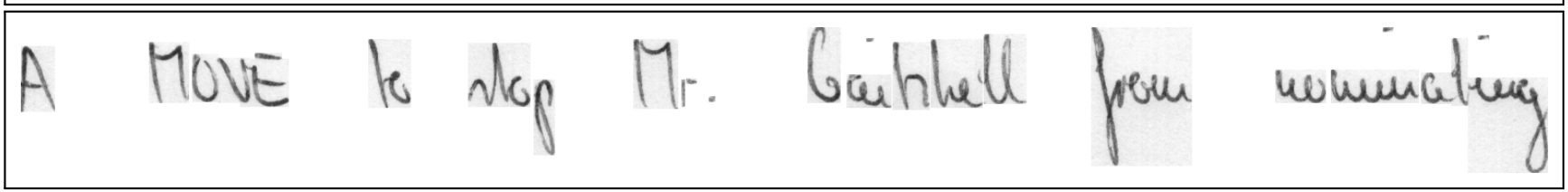
Why is it an Important Problem?

Handwritten Text Recognition









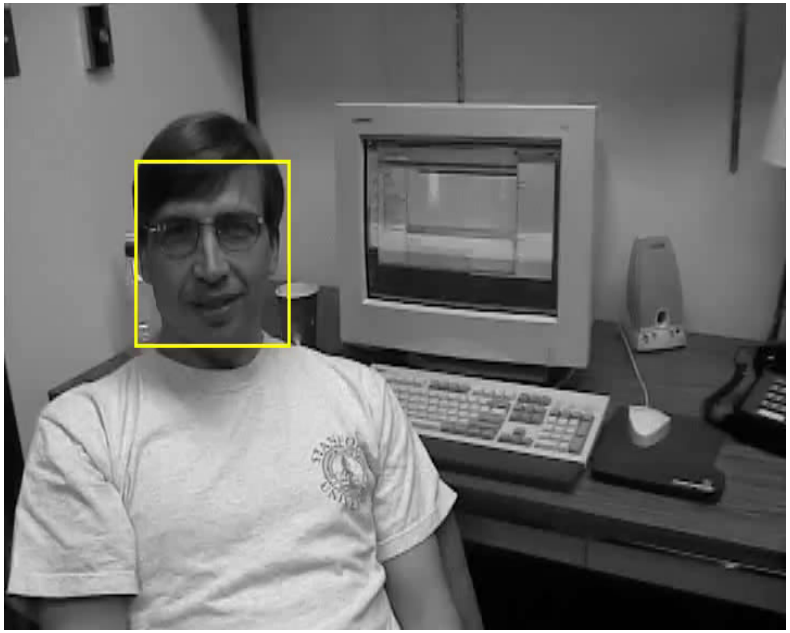
Challenges:

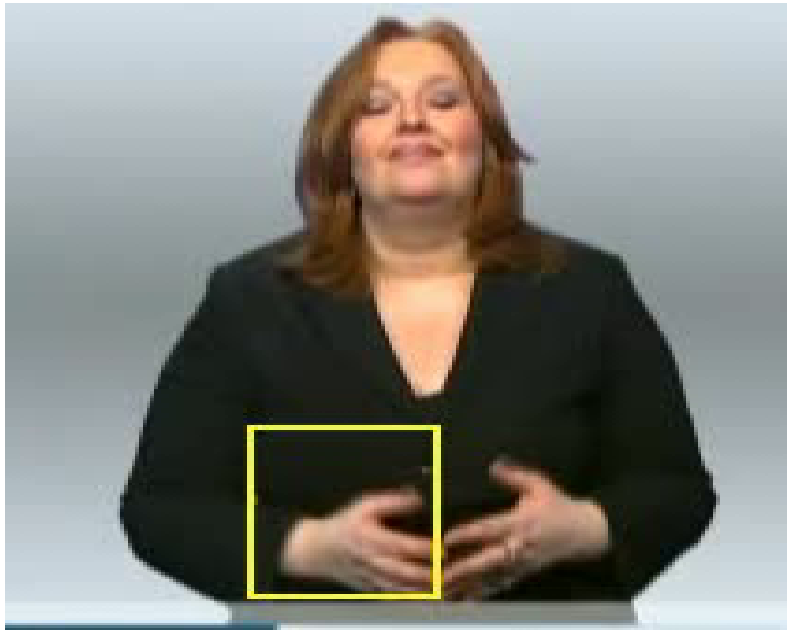
⇒ connected continuous handwritten texts

⇒ writer dependent handwriting styles

Why is it an Important Problem?

Head and Hand Tracking





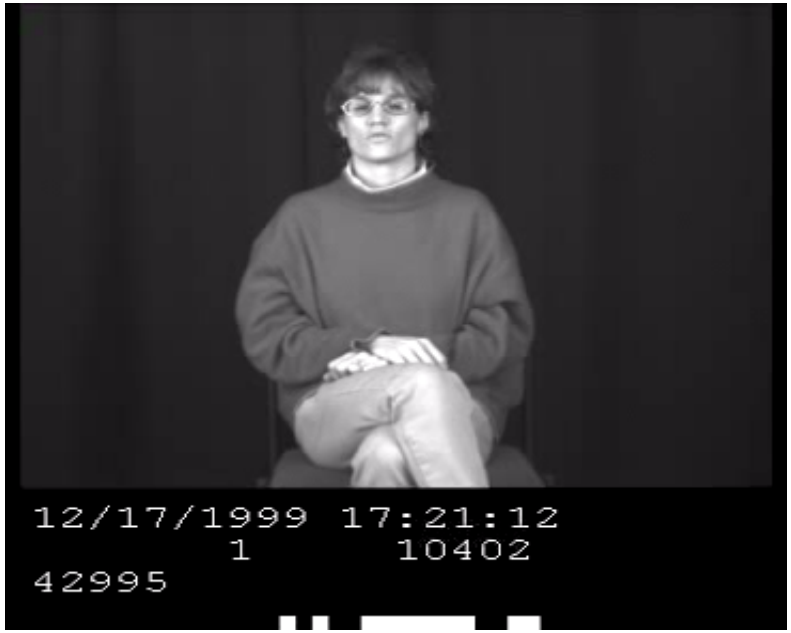
Challenges:

- ⇒ partially-occluded non-rigid objects
- ⇒ fast and abrupt movements

Why is it an Important Problem?

Introduction

Gesture and Sign Language Recognition



Challenges:

- ⇒ movement epenthesis and coarticulation effects
- ⇒ natural signed languages, i.e. national with local dialects

Scientific Goals

Introduction

Domains

- optical character recognition
- object tracking
- automatic sign language recognition

Some Questions

- which concepts and ideas can be adopted from ASR?
- can we tackle all domains within a unique framework?

Underlying Principles

- avoid early and local decisions
- quantitatively evaluate the improvements

Automatic Sign Language Recognition

RWTH AACHEN UNIVERSITY

Introduction

Automatic Sign Language Recognition

What Features do we need?

- manual: hand motion / form / orientation / location
- non-manual: mimic, eye gaze, body/head orientation

$\Rightarrow$ should be extracted from input signal

Different Approaches / Assumptions

- special hardware, computer vision, environment, ...

$\Rightarrow$ vision-based approaches do not restrict the way of signing

Introduction

Automatic Sign Language Recognition

Problems in many State-of-the-Art Approaches

- too controlled conditions
- most systems: person dependent, recognition of isolated signs
- lack of data, no publicly available corpora

Goals: Follow Approaches Similar to Speech Recognition

- recognition of continuous sign language
- training with sentences (unknown word boundaries)
- multi-person / person-independent training and recognition
- cope with dialects
- "large" datasets

$\Rightarrow$ extend RWTH-ASR large vocabulary speech recognition system

Signed-Language-to-Spoken-Language

Automatic Sign Language Recognition

Recognition: Sign-to-Text (Video $\Rightarrow$ Glosses)

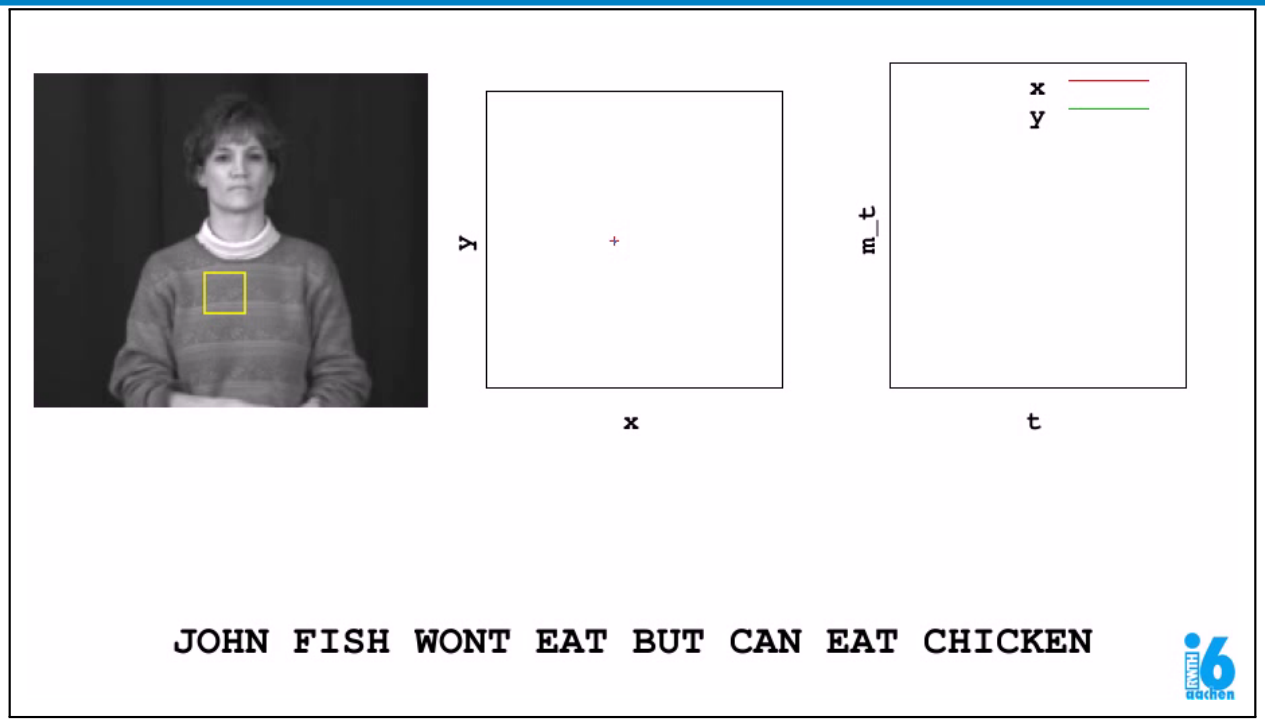
Translation: Text-to-Text (Glosses $\Rightarrow$ Text)

Synthesis: Text-to-Speech (Text $\Rightarrow$ Audio)

Signed-Language-to-Spoken-Language

Automatic Sign Language Recognition

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



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

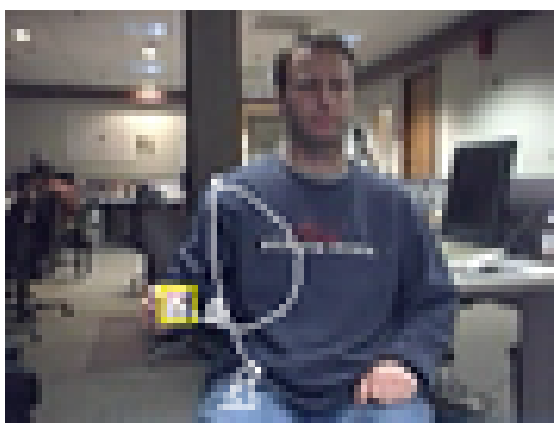
Synthesis: Text-to-Speech (Text ⇒ Audio)

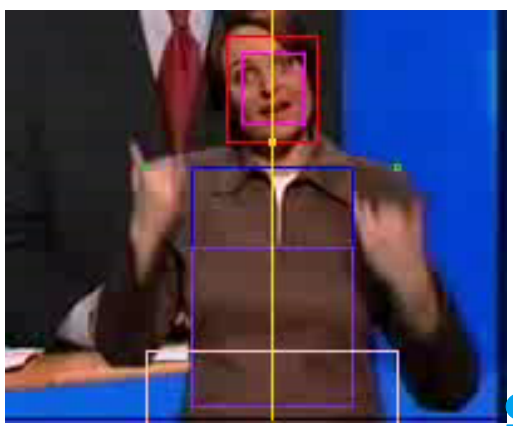
Multi-Purpose Object Tracking by Dynamic Programming Tracking

- model-free tracking approach based on dynamic programming

  [Dreuw & Deselaers+ 06], IEEE FG

  $\Rightarrow$ 2 steps: score calc. & traceback (full temporal context)





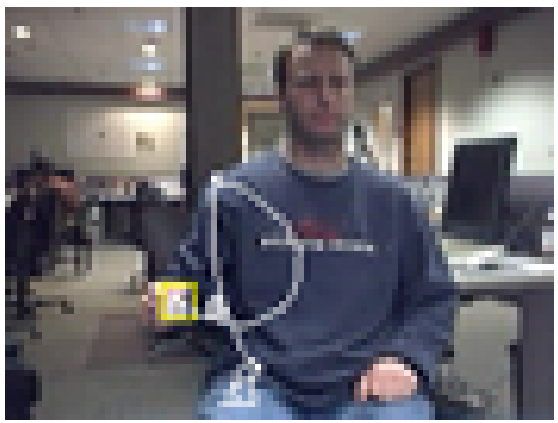
Multi-Purpose Object Tracking by Dynamic Programming Tracking

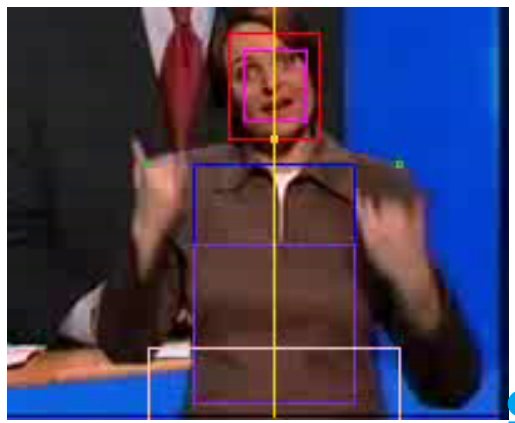
- model-free tracking approach based on dynamic programming

  [Dreuw & Deselaers+ 06], IEEE FG

  $\Rightarrow$ 2 steps: score calc. & traceback (full temporal context)
- common problem in sign language recognition:
  - tracking as pre-processing
  - path only optimized w.r.t. a tracking criterion
    (e.g. motion, color, etc.)

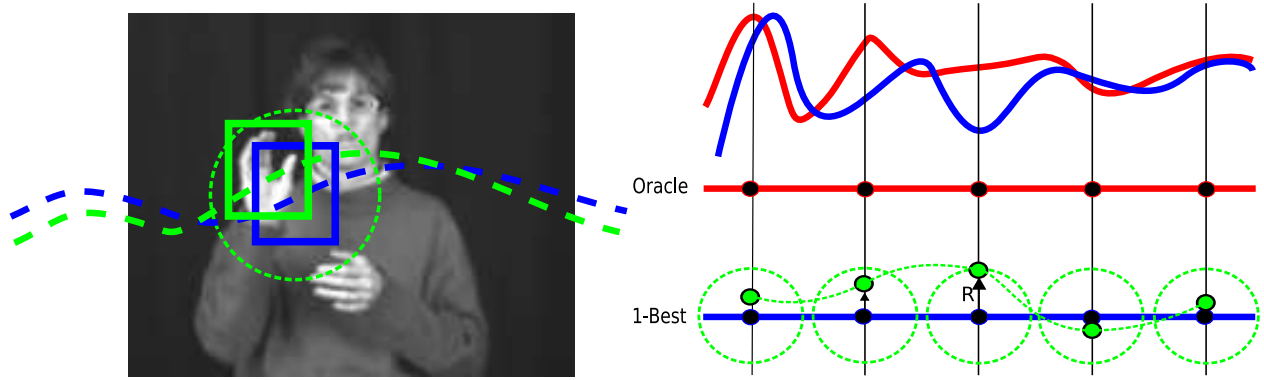
  $\Rightarrow$ early tracking decisions can lead to recognition errors





Tracking Extension for Sign Language Processing

- model-based tracking path adaptation [Dreuw & Forster+ 08], IEEE FG
  - consider positions around tracking path $\boldsymbol{u_1^T}$ within range $\boldsymbol{R}$
  - simultaneous tracking u_1^T and word sequence w_1^N optimization



Feature Extraction and Modeling

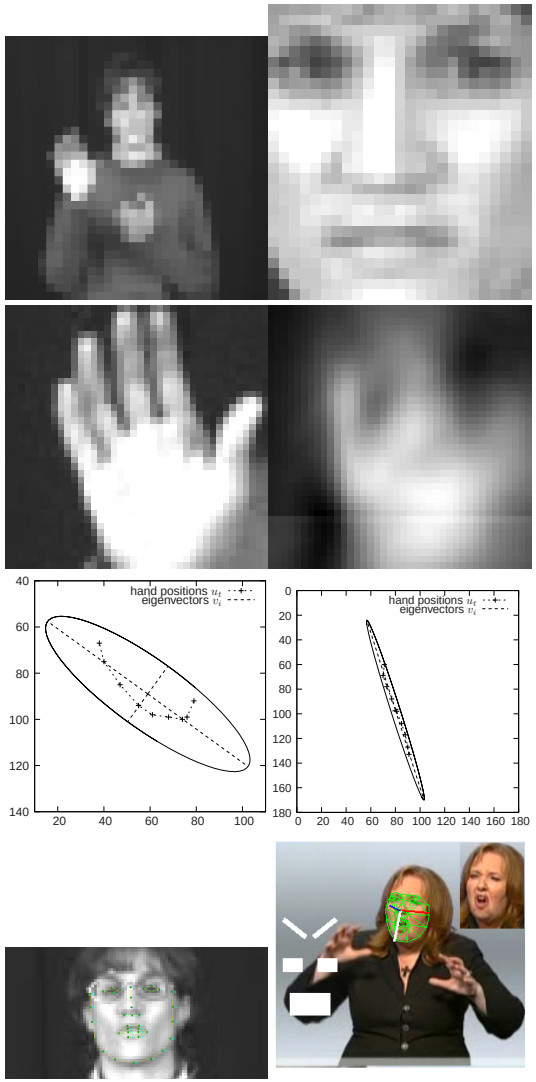
Automatic Sign Language Recognition

Features

- PCA-Frame
- PCA-Hand-Patch
- Hand position and trajectory
- Mean-Face
- AAM-based facial features
- ...

Modeling

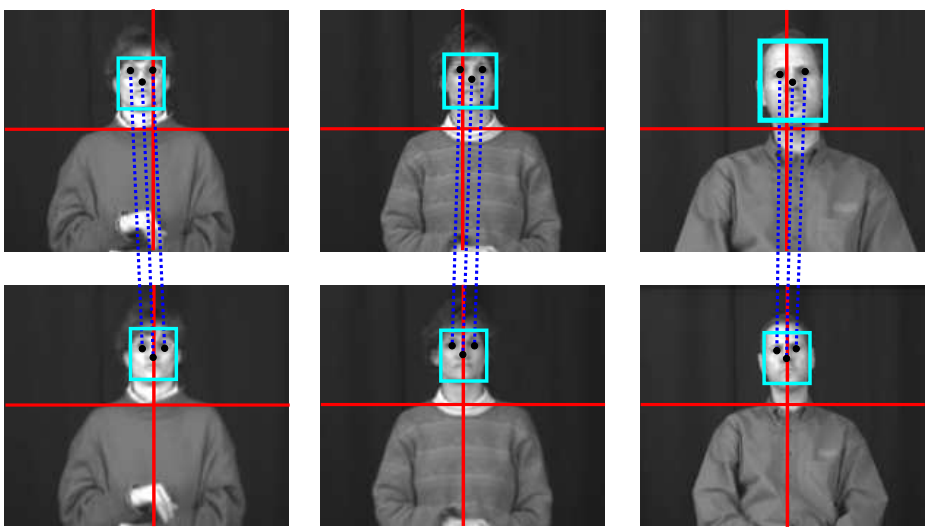
- Gaussian Mixture Models
- whole-word models
- adaptive lengths



Visual Speaker Alignment (VSA)

- appearance-based features: models are too speaker dependent

$\Rightarrow$ visually align speakers: scale and speaker independent features



Feature Extraction and Modeling

Automatic Sign Language Recognition

Visual Speaker Alignment (VSA)

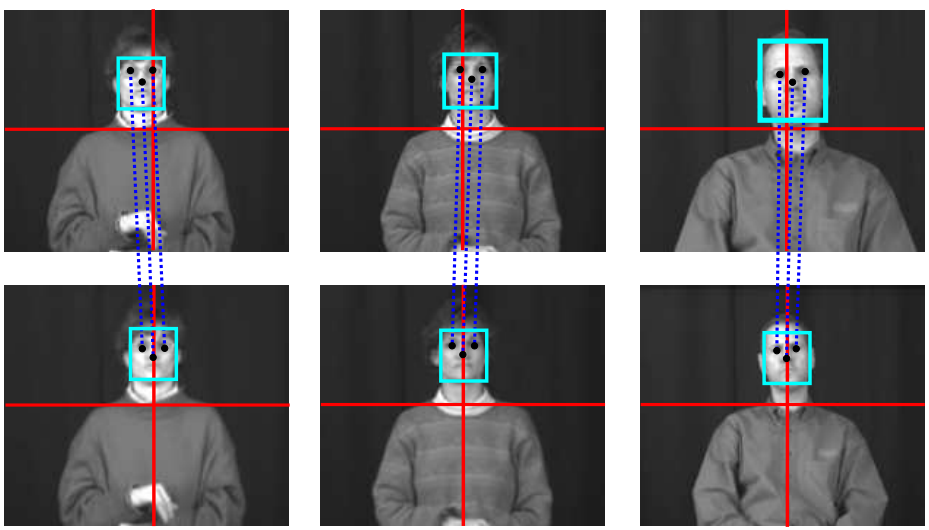
- appearance-based features: models are too speaker dependent

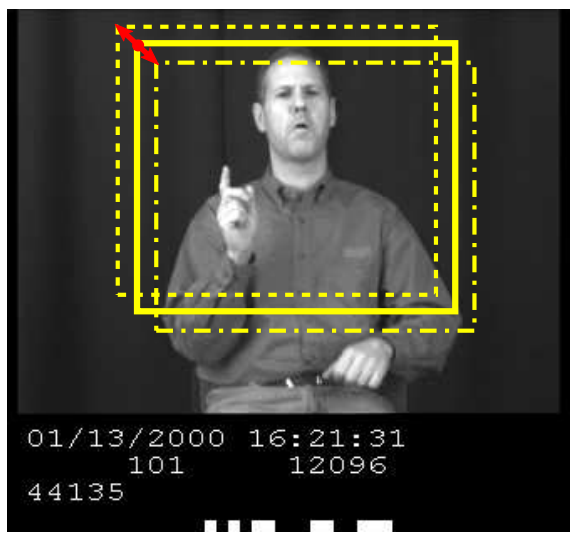
$\Rightarrow$ visually align speakers: scale and speaker independent features

Virtual Training Samples (VTS)

- lack of data problem

$\Rightarrow$ use virtual training samples





Experimental Results

Automatic Sign Language Recognition

RWTH-BOSTON-104 Database

- Corpus statistics

	training set	test set
# sentences	161	40
# running words	710	178
# frames	12422	3324
vocabulary size	103	65
# singletons	27	9
# out of vocabulary (OOV)	-	1
3-gram LM PP		4.7
signers		3

Experimental Results

Automatic Sign Language Recognition

Features / Rescoring	WER [%]			
	Baseline	VSA	VTS	VSA+VTS
Frame 32×32	38.76	33.15	27.53	24.72
PCA-Frame (200)	30.34	27.53	19.10	17.98

Experimental Results

Automatic Sign Language Recognition

Features / Rescoring	WER [%]			
	Baseline	VSA	VTS	VSA+VTS
Frame 32×32	38.76	33.15	27.53	24.72
PCA-Frame (200)	30.34	27.53	19.10	17.98
Hand (32×32)	45.51	34.83	25.28	31.46
+ distortion ($\boldsymbol{R = 10}$)	44.94	30.53	17.42	21.35
+ $\boldsymbol{\delta}$-penalty	35.96	28.65	18.54	20.79

Experimental Results

Automatic Sign Language Recognition

Features / Rescoring	WER [%]			
	Baseline	VSA	VTS	VSA+VTS
Frame 32×32	38.76	33.15	27.53	24.72
PCA-Frame (200)	30.34	27.53	19.10	17.98
Hand (32×32)	45.51	34.83	25.28	31.46
+ distortion ($\boldsymbol{R = 10}$)	44.94	30.53	17.42	21.35
+ $\boldsymbol{\delta}$-penalty	35.96	28.65	18.54	20.79
PCA-Hand (70)	44.94	34.27	15.73	26.97
+ distortion ($\boldsymbol{R = 10}$)	56.74	34.83	14.61	12.92
+ $\boldsymbol{\delta}$-penalty	32.58	24.16	11.24	12.92

$\Rightarrow$ model-based tracking adaptation strongly improves the results

$\Rightarrow$ effects of VSA and VTS are cumulative

Optical Character Recognition

RWTH AACHEN UNIVERSITY

Overview

Optical Character Recognition

Terminology

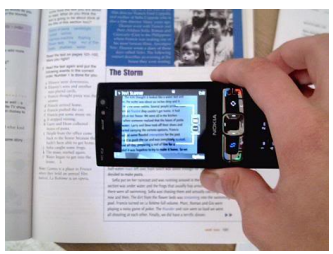
- OCR = optical character recognition (machine printed)
- ICR = intelligent character recognition (handwritten)

Common Requirements

- "flat" scans ⇒ line segments for recognition
- preprocessing: physical/logical layout analysis

Applications





State-of-the-Art

Optical Character Recognition

Commercial OCR Applications

- Novodynamics, Sakhrsoft, LEADTOOLS, ...
- OmniPage Pro 17, FineReader 10 Pro, ReadIRIS Pro 12

Freeware

- Google Docs OCR, free-ocr.com, ocrterminal.com, weocr, ...

Open Source Systems

- OCRopus, Tesseract-OCR v3.00, GOCR, OCRad
- HTK, RWTH OCR

$\Rightarrow$ comparison of systems/approaches is difficult

State-of-the-Art

Optical Character Recognition

Commercial OCR Applications

- Novodynamics, Sakhrsoft, LEADTOOLS, ...
- OmniPage Pro 17, FineReader 10 Pro, ReadIRIS Pro 12

Freeware

- Google Docs OCR, free-ocr.com, ocrterminal.com, weocr, ...

Open Source Systems

- OCRopus, Tesseract-OCR v3.00, GOCR, OCRad
- HTK, RWTH OCR

$\Rightarrow$ comparison of systems/approaches is difficult

$\Rightarrow$ support e.g. Arabic scripts

$\Rightarrow$ our goal: single framework, broad range of scripts/languages

State-of-the-Art

Optical Character Recognition

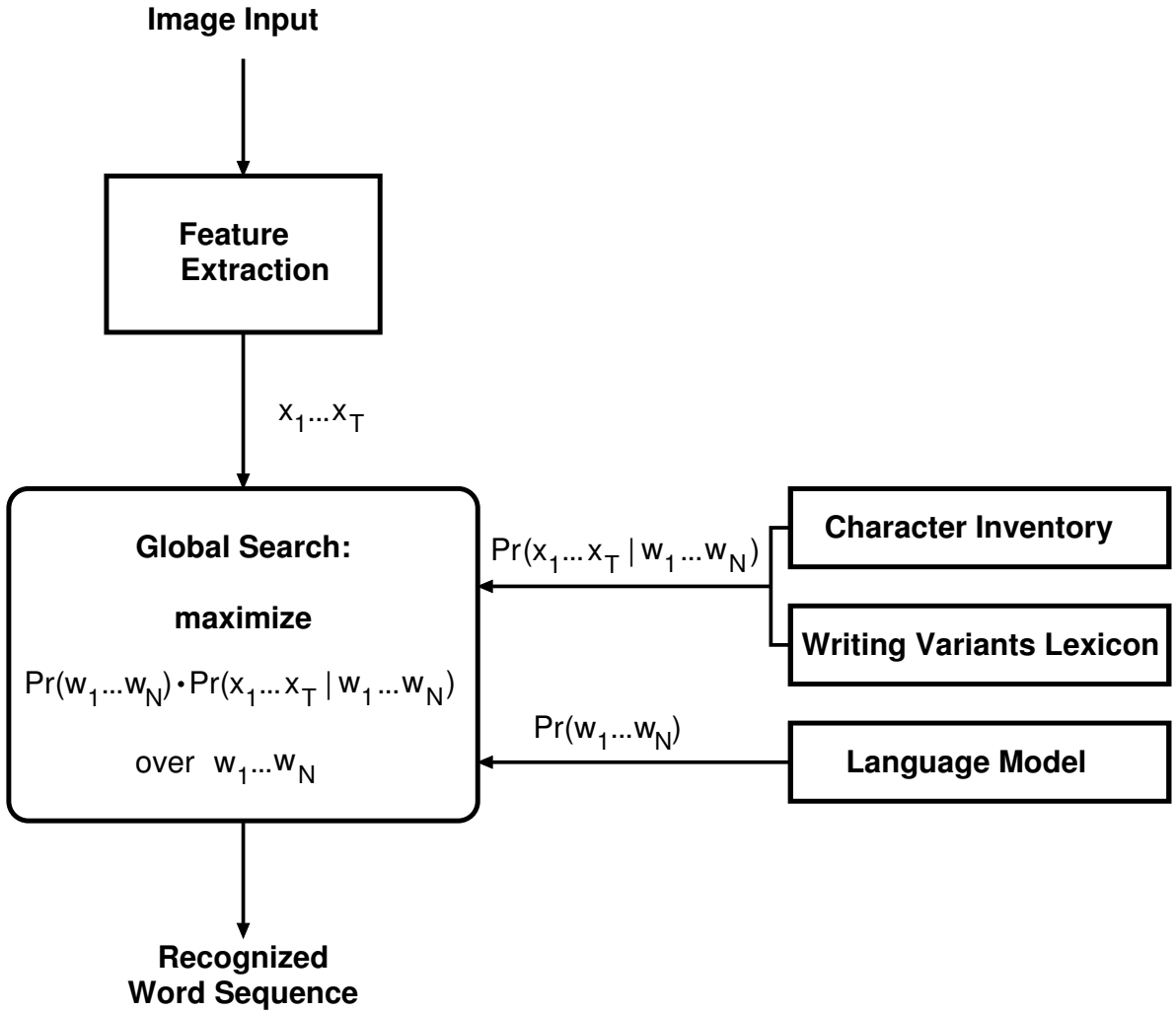
Research - Companies

- [Smith & Antonova$^+$ 09], Google
  Tesseract for Multilingual OCR, ICDAR 2009
- [Saleem & Cao$^+$ 09], BBN Technologies
  The BBN Byblos System, ICDAR 2009
- [Kermorvant & Menasri$^+$ 10], A2iA
  MLP/HMM-based Handwriting Recognition, ICFHR 2010

Research - Universities

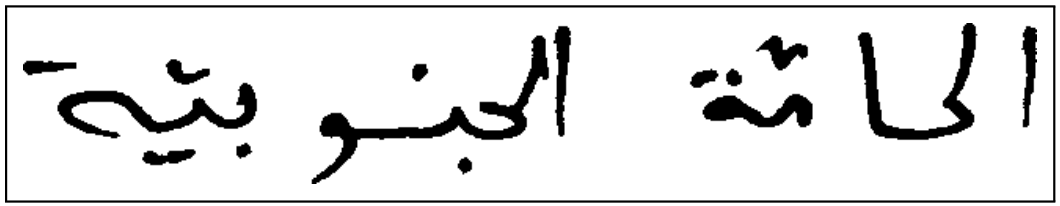
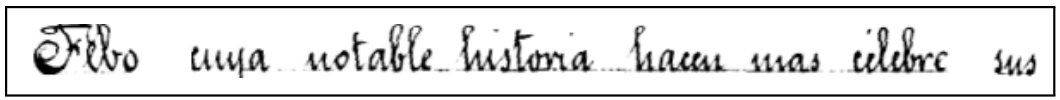
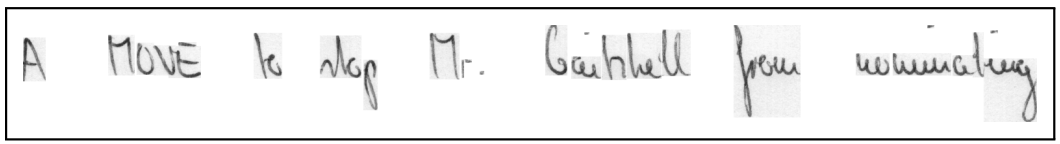
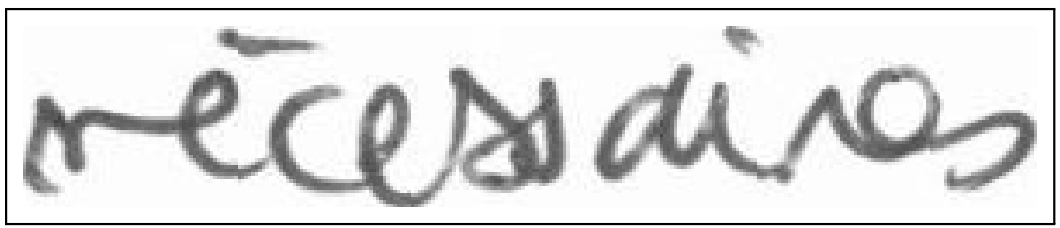
- [Bertolami & Bunke 08b], IAM
  HMM-based Handwriting Recognition, PR 2008
- [Graves & Liwicki$^+$ 09], TUM
  RNN/CTC-based Handwriting Recognition, PAMI 2009
- [Espana-Boquera & Castro-Bleda$^+$ 11], UPV
  ANN/HMM-based Handwriting Recognition, PAMI 2011

Recognition System



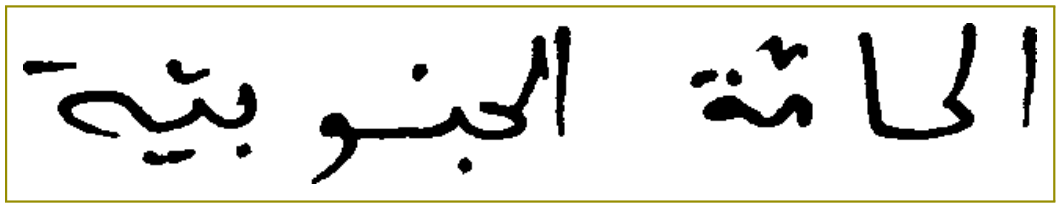
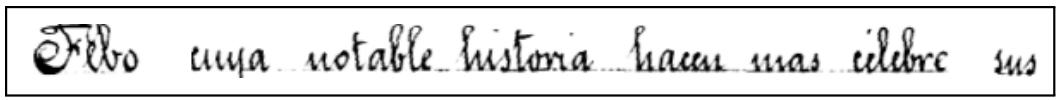
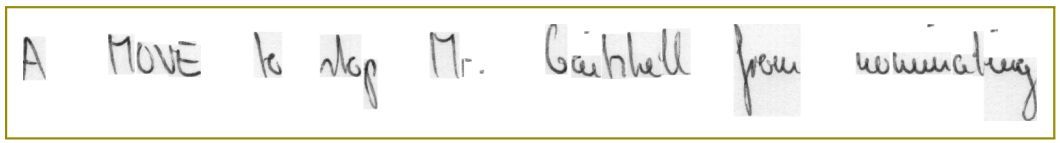
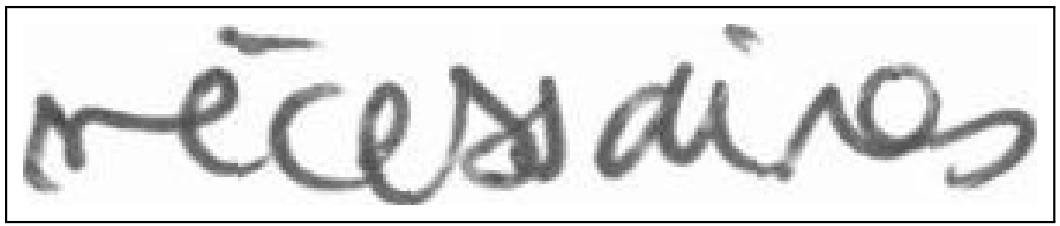
What Scripts can RWTH OCR Recognize?

Optical Character Recognition

Language	Database	Example
Arabic	IfN/ENIT	
Arabic	RAMP-N	
Catalan	GERMANA	
English	IAM	
French	RIMES	

What Scripts can RWTH OCR Recognize?

Optical Character Recognition

Language	Database	Example
Arabic	IfN/ENIT	
Arabic	RAMP-N	
Catalan	GERMANA	
English	IAM	
French	RIMES	

Features

Optical Character Recognition

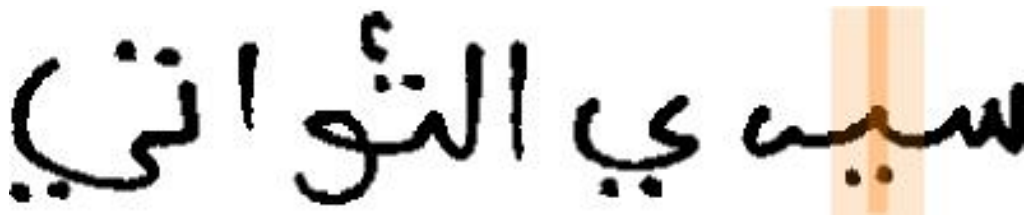
Preprocessing - Segment Normalization

- Latin handwriting: color, slant, and height normalization
- Arabic handwriting: no preprocesing!
- machine-print: skew

$\Rightarrow$ concept: avoid early decisions, focus on modeling

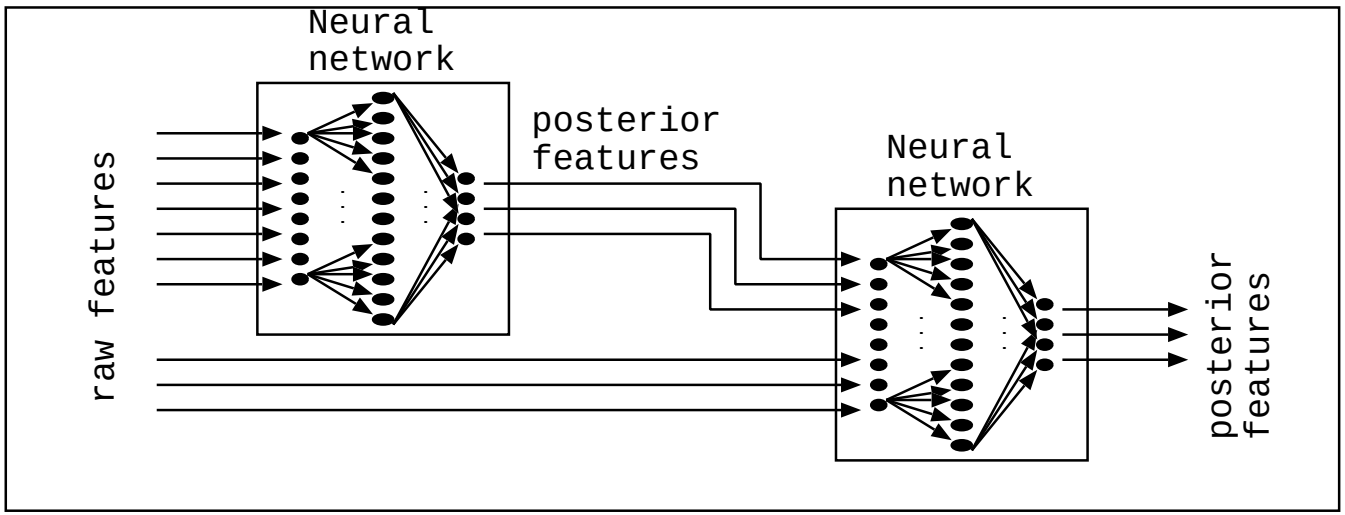
Appearance-Based

- sliding window, PCA reduction
- typically: large context-window with maximum overlap



Multi-Layer Perceptron (MLP)

- non-linear context modeling
- hierarchical neural network structures
- typically: 2 cascades, 1 hidden layer, large context-windows



$\Rightarrow$ RAW and TRAP-DCT posterior features

$\Rightarrow$ can be used for hybrid or tandem approaches

Glyph Modeling

Optical Character Recognition

Bayes' Decision Rule and HMMs

$$x_1^T \rightarrow \hat{w}_1^N(x_1^T) = \arg\max_{w_1^N} \left\{ p(w_1^N)\ p(x_1^T|w_1^N) \right\}$$

$$p(x_1^T|w_1^N) = \max_{[s_1^T]} \left\{ \prod_{t=1}^{T} p(x_t|s_t, w_1^N)\ p(s_t|s_{t-1}) \right\}$$

Gaussian HMMs

- Gaussian mixture models as emissions
- left-to-right topology with skip transitions

Bayes' Decision Rule and HMMs

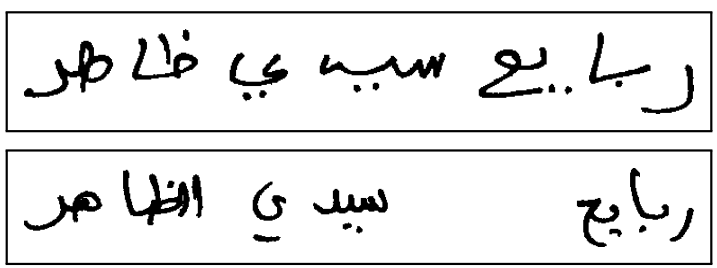
$$x_1^T \rightarrow \hat{w}_1^N(x_1^T) = \arg\max_{w_1^N} \left\{ p(w_1^N)\ p(x_1^T|w_1^N) \right\}$$

$$p(x_1^T|w_1^N) = \max_{[s_1^T]} \left\{ \prod_{t=1}^{T} p(x_t|s_t, w_1^N)\ p(s_t|s_{t-1}) \right\}$$

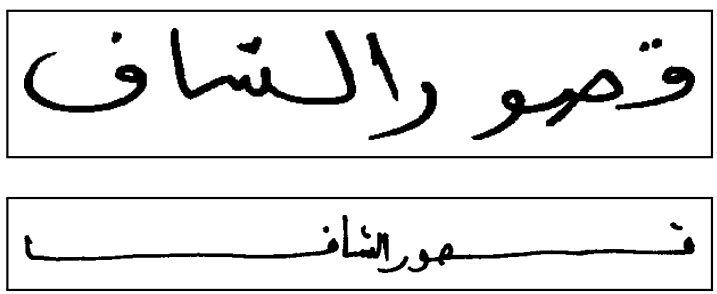
Gaussian HMMs

- Gaussian mixture models as emissions
- left-to-right topology with skip transitions

$\Rightarrow$ important in Arabic handwriting:



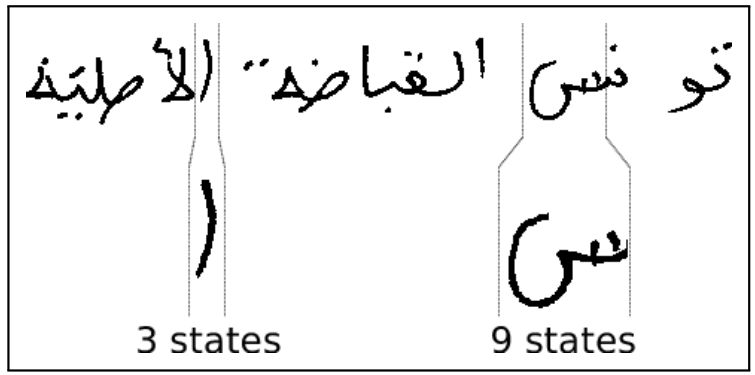
(a) white-spaces



(b) state-transitions

Glyph Dependent Lengths (GDL) for GHMMs

- wide complex characters $\Rightarrow$ more HMM states

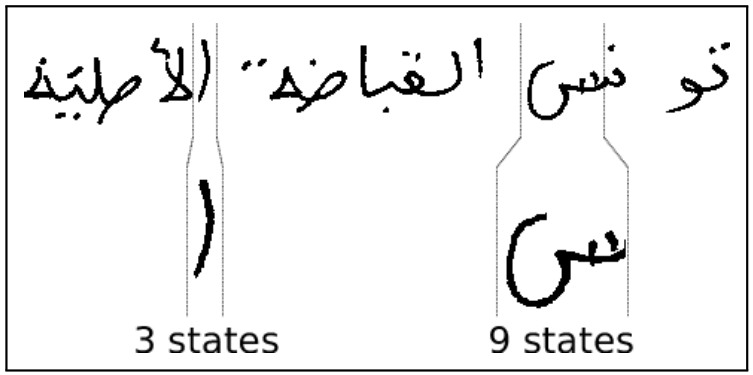


- unsupervised iterative approach: update #states S_c for each glyph model c by alignment and frequency counts

$$S_c = \frac{\mathsf{N}(x, c)}{\mathsf{N}(c)} \cdot \alpha$$

Glyph Dependent Lengths (GDL) for GHMMs

- wide complex characters $\Rightarrow$ more HMM states



- unsupervised iterative approach: update #states S_c for each glyph model c by alignment and frequency counts

$$S_c = \frac{\mathsf{N}(x, c)}{\mathsf{N}(c)} \cdot \alpha$$

$\Rightarrow$ important in Arabic handwriting [Dreuw & Jonas+ 08], ICPR

$\Rightarrow$ less important: Arabic machine-print, w/ preprocessing

Hybrid MLP/HMM

- approximate the observation probabilities of an HMM

$$p(x_t|s_t) = \frac{p(s_t|x_t)}{p(s_t)}$$

  - $p(s_t|x_t)$ realized as MLP posterior feature stream (offline)
  - $p(s_t)$ prior provided by previously trained model

Tandem MLP-GHMM

- estimate using log-PCA reduced MLP posterior probabilities

$$x'_t = \phi(\log p(s_t|x_t))$$

Glyph Modeling

Optical Character Recognition

Hybrid MLP/HMM

- approximate the observation probabilities of an HMM

$$p(x_t|s_t) = \frac{p(s_t|x_t)}{p(s_t)}$$

  - $p(s_t|x_t)$ realized as MLP posterior feature stream (offline)
  - $p(s_t)$ prior provided by previously trained model

Tandem MLP-GHMM

- estimate using log-PCA reduced MLP posterior probabilities

$$x'_t = \phi(\log p(s_t|x_t))$$

$\Rightarrow$ important is initial MLP alignment [Dreuw & Doetsch+ 11], IEEE ICIP

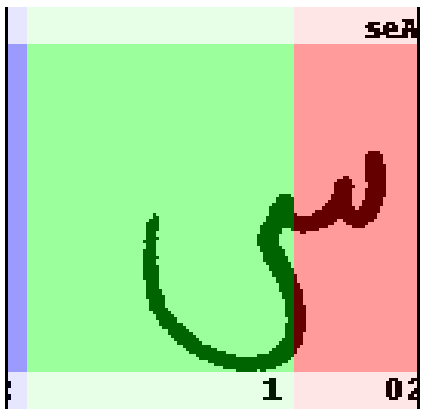
Writing Variants Lexicon

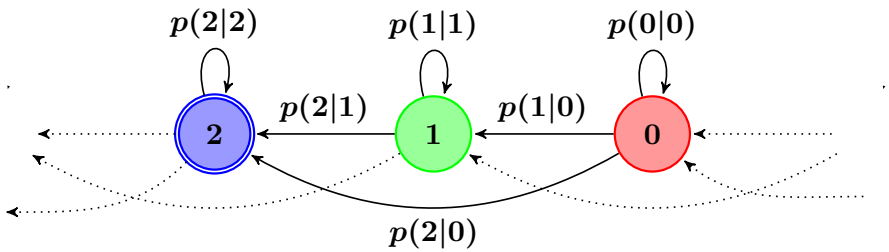
Optical Character Recognition

Arabic Scripts

- ligatures and diacritics ⇒ multiple writing variants
- PAWs: other approaches had difficulties ⇒ explicit modeling

Example





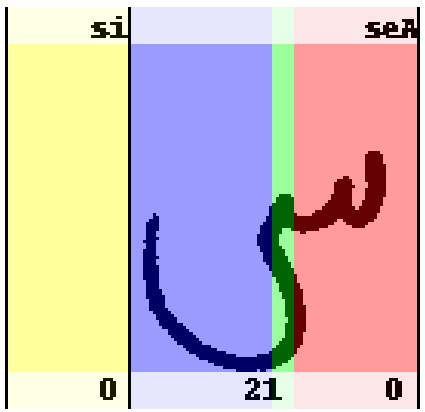
Writing Variants Lexicon

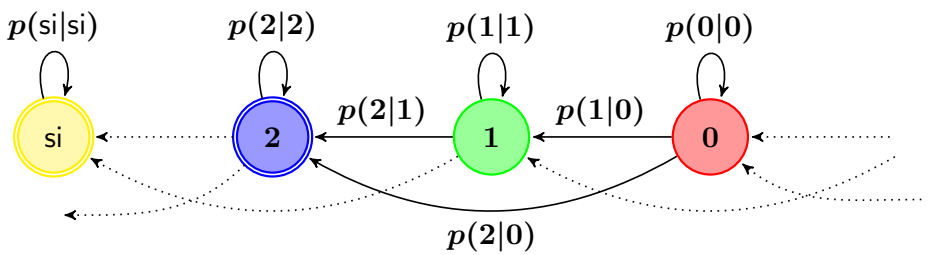
Optical Character Recognition

Arabic Scripts

- ligatures and diacritics ⇒ multiple writing variants
- PAWs: other approaches had difficulties ⇒ explicit modeling

Example





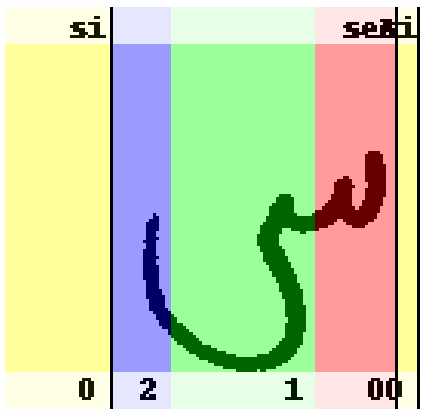
Writing Variants Lexicon

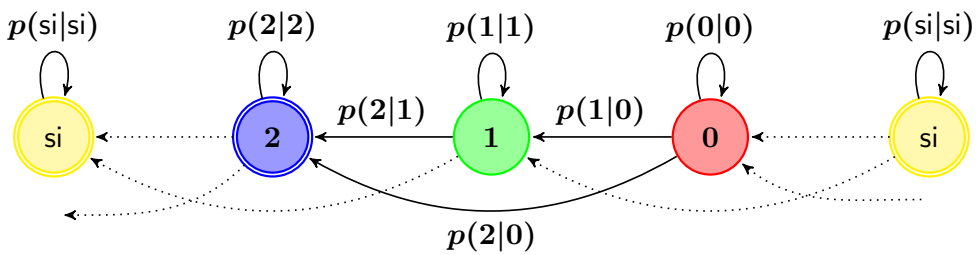
Optical Character Recognition

Arabic Scripts

- ligatures and diacritics $\Rightarrow$ multiple writing variants
- PAWs: other approaches had difficulties $\Rightarrow$ explicit modeling

Example



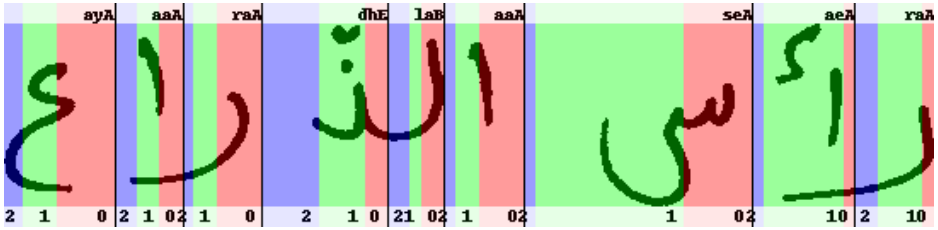


Writing Variants Lexicon

Optical Character Recognition

Arabic Scripts: Explicit White-Space Modeling

without

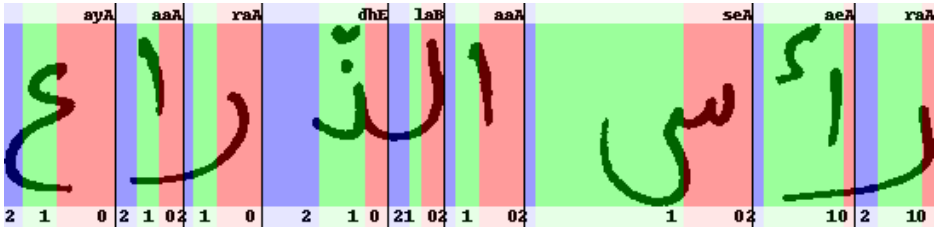


Writing Variants Lexicon

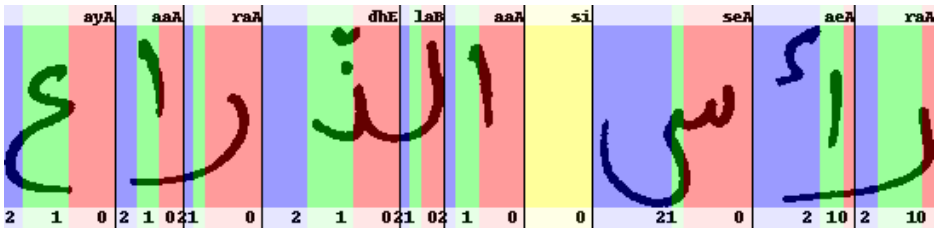
Optical Character Recognition

Arabic Scripts: Explicit White-Space Modeling

without



between compounds

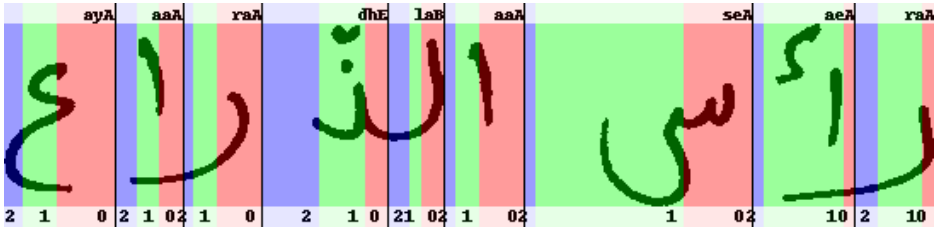


Writing Variants Lexicon

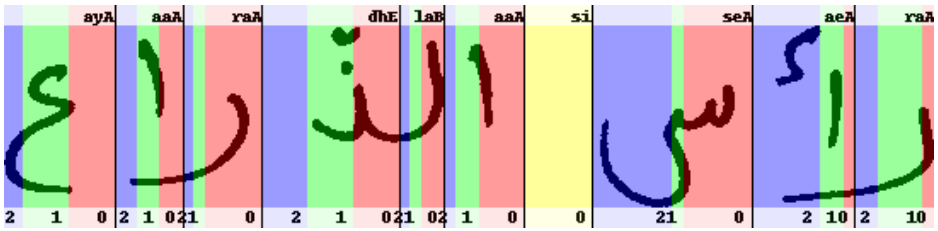
Optical Character Recognition

Arabic Scripts: Explicit White-Space Modeling

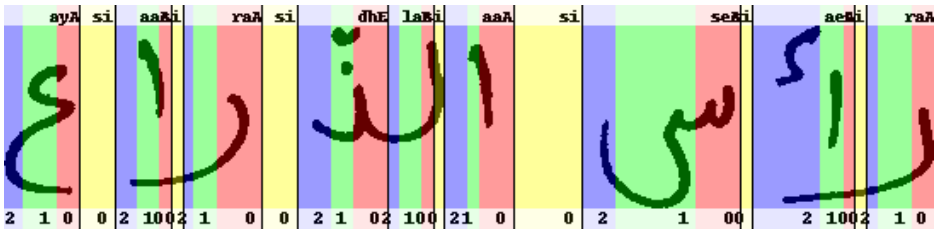
without



between compounds



between and within
(as writing variants)



⇒ important in Arabic handwriting [Dreuw & Jonas+ 08], ICPR

⇒ less important: Arabic machine-print

Training and Decoding Architectures

Optical Character Recognition

Training

- Maximum Likelihood (ML)
- Writer Adaptive Training (WAT) [Dreuw & Rybach+ 09], ICDAR
- Discriminative training criteria (M-MMI/M-MPE)
- Tandem (MLP-GHMM)

Decoding

- 1-pass
  - GHMM / Tandem MLP-GHMM model
  - Hybrid MLP/HMM
- 2-pass
  - Writer Adaptation
  - Unsupervised Confidence-based Discriminative Training

▸ Details

Training and Decoding Architectures

Optical Character Recognition

Training

- Maximum Likelihood (ML)
- Writer Adaptive Training (WAT) [Dreuw & Rybach+ 09], ICDAR
- Discriminative training criteria (M-MMI/M-MPE)
- Tandem (MLP-GHMM)

Decoding

- 1-pass
  - GHMM / Tandem MLP-GHMM model
  - Hybrid MLP/HMM
- 2-pass
  - Writer Adaptation
  - Unsupervised Confidence-based Discriminative Training

▸ Details

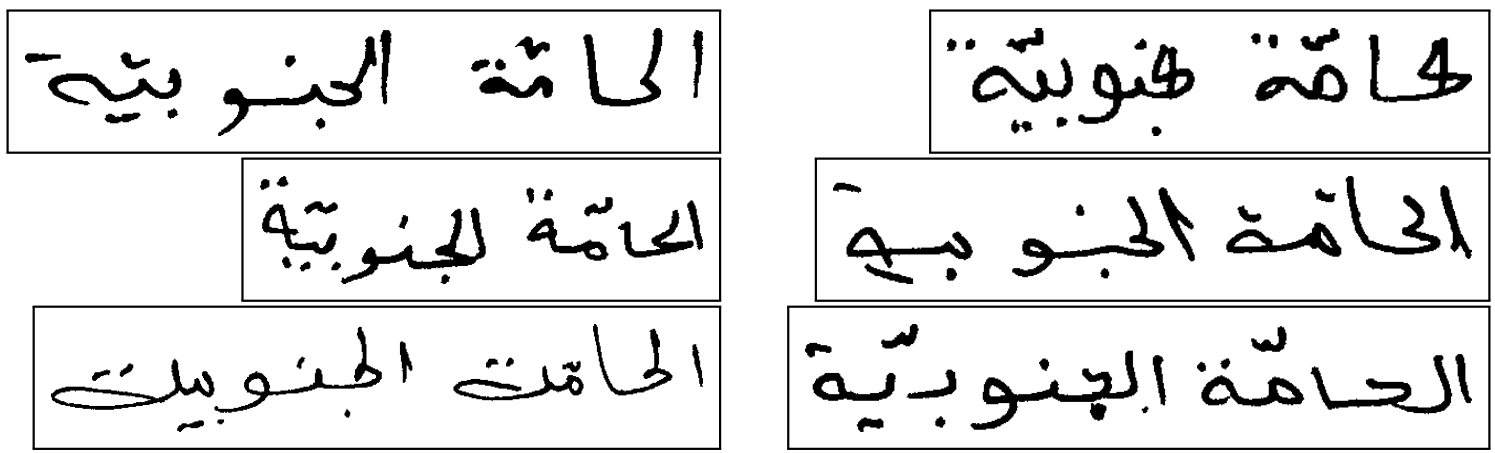
Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Statistics

- 937 Tunisian city names
- 32492 handwritten Arabic words, 916 writers, several sets
- database is used by more than 60 groups all over the world

Example (same city name)



Baseline System

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Comparisons and Progress: ICDAR / ICFHR Competitions

external evaluations on unknown sets f and s [Märgner & Abed 10]

Year	Group (Approach)	set-e	set-f	set-s
ICDAR 2009	MDLSTM (RNN/CTC)	-	**6.6**	18.9
	A2iA (GHMM & MLP/HMM)	-	10.6	23.3
	RWTH OCR (x16, GHMM, M-MMI)	15.4	14.5	28.7
	RWTH OCR (x16, GHMM, M-MMI-conf)	14.6	14.3	27.5

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Comparisons and Progress: ICDAR / ICFHR Competitions

external evaluations on unknown sets f and s [Märgner & Abed 10]

Year	Group (Approach)	set-e	set-f	set-s
ICDAR 2009	MDLSTM (RNN/CTC)	-	**6.6**	18.9
	A2iA (GHMM & MLP/HMM)	-	10.6	23.3
	RWTH OCR (x16, GHMM, M-MMI)	15.4	14.5	28.7
	RWTH OCR (x16, GHMM, M-MMI-conf)	14.6	14.3	27.5
ICFHR 2010	UPV PRHLT (HMM)	6.2	7.8	**15.4**
	RWTH OCR (x16, MLP-GHMM, M-MMI)	7.3	9.1	18.9
	CUBS-AMA (HMM)	-	19.7	32.1

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Comparisons and Progress: ICDAR / ICFHR Competitions

external evaluations on unknown sets f and s [Märgner & Abed 10]

Year	Group (Approach)	set-e	set-f	set-s
ICDAR 2009	MDLSTM (RNN/CTC)	-	**6.6**	18.9
	A2iA (GHMM & MLP/HMM)	-	10.6	23.3
	RWTH OCR (x16, GHMM, M-MMI)	15.4	14.5	28.7
	RWTH OCR (x16, GHMM, M-MMI-conf)	14.6	14.3	27.5
ICFHR 2010	UPV PRHLT (HMM)	6.2	7.8	**15.4**
	RWTH OCR (x16, MLP-GHMM, M-MMI)	7.3	9.1	18.9
	CUBS-AMA (HMM)	-	19.7	32.1
ICDAR 2011	RWTH OCR (x32, MLP-GHMM, ML)	**5.9**	7.8	15.5
	REGIM (HMM)	-	21.0	31.6
	JU-OCR (RF & Rules)	-	36.1	50.3

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

- English handwriting
- LM: Brown, Lancester-Oslo-Bergen, and Wellington corpora
- 50k lexicon, 3-gram LM

	Train	Devel	Eval	LM
words	53.8k	8.7k	25.4k	3.3M
chars	219.7k	31.7k	96.6k	13.8M
lines	6.1k	0.9k	2.7k	164k
writers	283	57	162	-
OOV rate	1.07%	3.94%	3.42%	1.87%

▸ MMI/MPE Results ▸ Unsupervised Results

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9
+ M-MMI	25.8	31.6

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9
+ M-MMI	25.8	31.6
+M-MMI-conf	23.7	29.0

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9
+ M-MMI	25.8	31.6
+M-MMI-conf	23.7	29.0
+ M-MPE	24.3	30.0

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9
+ M-MMI	25.8	31.6
+M-MMI-conf	23.7	29.0
+ M-MPE	24.3	30.0
+ M-MPE-conf	23.7	29.2

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9
+ M-MMI	25.8	31.6
+M-MMI-conf	23.7	29.0
+ M-MPE	24.3	30.0
+ M-MPE-conf	23.7	29.2
MLP/HMM	31.2	36.9

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9
+ M-MMI	25.8	31.6
+M-MMI-conf	23.7	29.0
+ M-MPE	24.3	30.0
+ M-MPE-conf	23.7	29.2
MLP/HMM	31.2	36.9
MLP-GHMM	25.7	32.9
+ M-MMI	23.5	30.1
+ M-MPE	22.7	28.8

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9
+ M-MMI	25.8	31.6
+M-MMI-conf	23.7	29.0
+ M-MPE	24.3	30.0
+ M-MPE-conf	23.7	29.2
MLP/HMM	31.2	36.9
MLP-GHMM	25.7	32.9
+ M-MMI	23.5	30.1
+ M-MPE	22.7	28.8
[Bertolami & Bunke 08a] (GHMMs)	26.8	32.8
[Graves & Liwicki^{+} 09] (LSTM/CTC)	-	25.9
[Espana-Boquera & Castro-Bleda^{+} 11] (MLPs/HMM)	19.0	22.4

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]	
	Devel	Eval
GHMM, ML baseline	31.9	38.9
+ M-MMI	25.8	31.6
+M-MMI-conf	23.7	29.0
+ M-MPE	24.3	30.0
+ M-MPE-conf	23.7	29.2
MLP/HMM	31.2	36.9
MLP-GHMM	25.7	32.9
+ M-MMI	23.5	30.1
+ M-MPE	22.7	28.8
[Bertolami & Bunke 08a] (GHMMs)	26.8	32.8
[Graves & Liwicki^{+} 09] (LSTM/CTC)	-	25.9
[Espana-Boquera & Castro-Bleda^{+} 11] (MLPs/HMM)	19.0	22.4
[Doetsch 11] RWTH OCR (LSTM-GHMM, M-MPE)	**17.4**	**21.4**

Conclusions and Future Work

RWTH AACHEN UNIVERSITY

Conclusions

Conclusions and Future Work

Optical Character Recognition

- able to recognize handwritten and machine printed texts
- MLP and HMM can cope with horizontal variations/contexts
- neural network based features significant improvements
- excellent results, also at external evaluations

Object Tracking

- multi-purpose tracking framework (DPT)
- robust and smooth head and hand trajectories
- excellent results on various datasets of different visual complexity

Automatic Sign Language Recognition

- many similarities with ASR
- good temporal alignments & adequate features are crucial

Future Work

Conclusions and Future Work

Features, Visual Modeling, Training, LMs, ...

- "intelligent" preprocessing
- robust / high-level features
- context modeling (e.g. CART)
- writer/font adaptive training
- joint optimization of neural networks and HMM
- unsupervised adaptation
- ...

Thank you for your attention

Philippe Dreuw

dreuw@cs.rwth-aachen.de

http://www.hltpr.rwth-aachen.de/~dreuw/

References I

[Bertolami & Bunke 08a] R. Bertolami, H. Bunke:

Hidden Markov model-based ensemble methods for offline handwritten text line recognition.

Pattern Recognition, Vol. 41, No. 11, pp. 3452–3460, Nov. 2008.

[Bertolami & Bunke 08b] R. Bertolami, H. Bunke:

HMM-based ensamble methods for offline handwritten text line recognition.

Pattern Recognition, Vol. 41, pp. 3452–3460, 2008.

[Doetsch 11] P. Doetsch:

Optimization of Hidden Markov Models and Neural Networks.

Master's thesis, RWTH Aachen University, Dec. 2011.

References II

[Dreuw & Deselaers+ 06] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, H. Ney:

Tracking Using Dynamic Programming for Appearance-Based Sign Language Recognition.

In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 293–298, Southampton, April 2006.

[Dreuw & Doetsch+ 11] P. Dreuw, P. Doetsch, C. Plahl, H. Ney:

Hierarchical Hybrid MLP/HMM or rather MLP Features for a Discriminatively Trained Gaussian HMM: A Comparison for Offline Handwriting Recognition.

In *IEEE International Conference on Image Processing (ICIP)*, pp. 1–4, Brussels, Belgium, Sept. 2011.

References III

[Dreuw & Forster+ 08] P. Dreuw, J. Forster, T. Deselaers, H. Ney:

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

In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, Amsterdam, The Netherlands, Sept. 2008.

[Dreuw & Heigold+ 11] P. Dreuw, G. Heigold, H. Ney:

Confidence and Margin-Based MMI/MPE Discriminative Training for Offline Handwriting Recognition.

International Journal on Document Analysis and Recognition (IJDAR), Vol. PP, pp. accepted for publication, March 2011.

DOI 10.1007/s10032-011-0160-x The final publication is available at www.springerlink.com.

References IV

[Dreuw & Jonas+ 08] P. Dreuw, S. Jonas, H. Ney:

White-Space Models for Offline Arabic Handwriting Recognition.

In *International Conference on Pattern Recognition (ICPR)*, pp. 1–4, Tampa, Florida, USA, Dec. 2008.

[Dreuw & Rybach+ 09] P. Dreuw, D. Rybach, C. Gollan, H. Ney:

Writer Adaptive Training and Writing Variant Model Refinement for Offline Arabic Handwriting Recognition.

In *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 21–25, Barcelona, Spain, July 2009.

[Dreuw & Stein+ 07] P. Dreuw, D. Stein, H. Ney:

Enhancing a Sign Language Translation System with Vision-Based Features.

In *Intl. Workshop on Gesture in HCI and Simulation 2007*, LNCS, pp. 18–19, Lisbon, Portugal, May 2007.

References V

[Espana-Boquera & Castro-Bleda$^+$ 11] S. Espana-Boquera, M. Castro-Bleda, J. Gorbe-Moya, F. Zamora-Martinez:

Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models.

IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 4, pp. 767–779, April 2011.

[Gollan & Bacchiani 08] C. Gollan, M. Bacchiani:

Confidence Scores for Acoustic Model Adaptation.

In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4289–4292, Las Vegas, NV, USA, April 2008.

[Graves & Liwicki$^+$ 09] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, J. Schmidhuber:

A Novel Connectionist System for Unconstrained Handwriting Recognition.

IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 5, pp. 855–868, May 2009.

References VI

[Heigold 10] G. Heigold:

A Log-Linear Discriminative Modeling Framework for Speech Recognition.

Ph.D. thesis, RWTH Aachen University, Aachen, Germany, June 2010.

[Heigold & Dreuw$^+$ 10] G. Heigold, P. Dreuw, S. Hahn, R. Schlüter, H. Ney:

Margin-Based Discriminative Training for String Recognition.

Journal of Selected Topics in Signal Processing - Statistical Learning Methods for Speech and Language Processing, Vol. 4, No. 6, pp. 917–925, Dec. 2010.

[Hermansky & Sharma 98] H. Hermansky, S. Sharma:

TRAPs - classifiers of temporal patterns.

In *International Conference on Spoken Language Processing (ICSLP)*, pp. 289–292, March 1998.

References VII

[Jonas 09] S. Jonas:

Improved Modeling in Handwriting Recognition.

Master's thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany, June 2009.

[Kermorvant & Menasri+ 10] C. Kermorvant, F. Menasri, A.L. Bianne, R. Al-Hajj, L. Likforman-Sulem, C. Mokbel:

The A2iA-Telecom ParisTech-UOB System for the ICDAR 2009 Handwriting Recognition Competition.

In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Kalkota, India, Nov. 2010.

[Märgner & Abed 10] V. Märgner, H.E. Abed:

ICFHR 2010 – Arabic Handwriting Recognition Competition.

In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 709–714, Nov. 2010.

References VIII

[Märgner & Abed 11] V. Märgner, H.E. Abed:

ICDAR 2011 – Arabic Handwriting Recognition Competition.

In *International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1444–1448, Sept. 2011.

[Povey & Woodland 02] D. Povey, P.C. Woodland:

Minimum phone error and I-smoothing for improved discriminative training.

In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 105–108, Orlando, FL, USA, May 2002.

[Saleem & Cao$^+$ 09] S. Saleem, H. Cao, K. Subramanian, M. Kamali, R. Prasad, P. Natarajan:

Improvements in BBN's HMM-based Offline Arabic Handwriting Recognition System.

In *International Conference on Document Analysis and Recognition*, pp. 773–777, Barcelona, Spain, July 2009.

References IX

[Smith & Antonova+ 09] R. Smith, D. Antonova, D.S. Lee:

Adapting the Tesseract open source OCR engine for multilingual OCR.

In *Proceedings of the International Workshop on Multilingual OCR*, pp. 1:1–1:8, New York, NY, USA, July 2009. ACM.

Appendix: Optical Character Recognition

RWTH AACHEN UNIVERSITY

State-of-the-Art

Optical Character Recognition

Competitions

- ICDAR / ICFHR:
  segmentation and recognition
  external evaluations
- DARPA MADCAT / NIST OpenHaRT:
  segmentation, recognition, and translation

Machine Printed Text Recognition

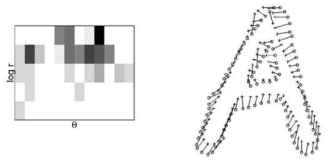
- many benchmark datasets available

$\Rightarrow$ Arabic?

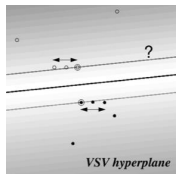
State-of-the-Art

Optical Character Recognition

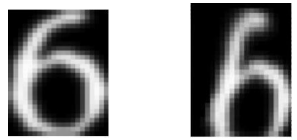
Belongie & Malik^{+} 2002 (Berkeley)
shape context matching



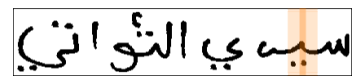
DeCoste & Schölkopf^{+} 2002 (CalTech / MPI)
invariant SVM



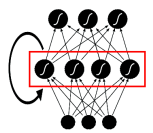
Simard & Steinkraus^{+} 2003 (MSR)
convolutional neural network



Schambach & Rottland^{+} 2008 (Siemens AG)
Natarajan et al. 2009 (BBN Technologies)
HMM



Graves et al. 2009 (TUM)
recurrent neural network

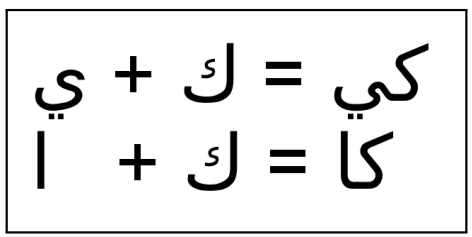


Arabic Writing System

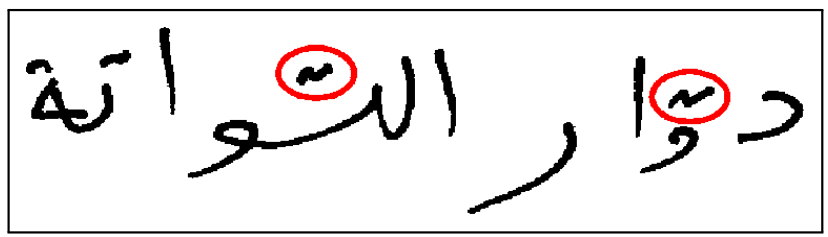
Optical Character Recognition

Arabic

- 28 base characters, up to 4 position dependent shapes
- ligatures, diacritics - optional in handwriting!
- Part of Arabic Word (PAW) as sub-words
- machine-print: cursive, shape usually not encoded!

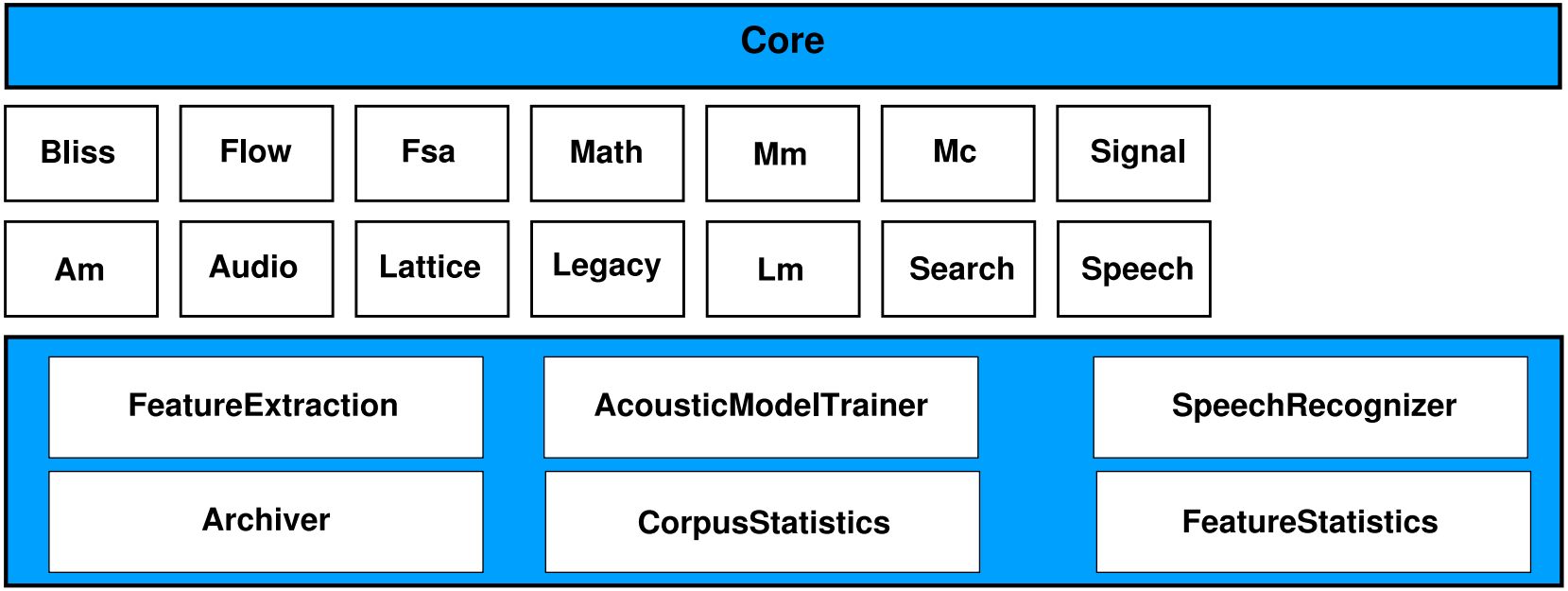


(a) Ligatures



(b) Diacritics

Software

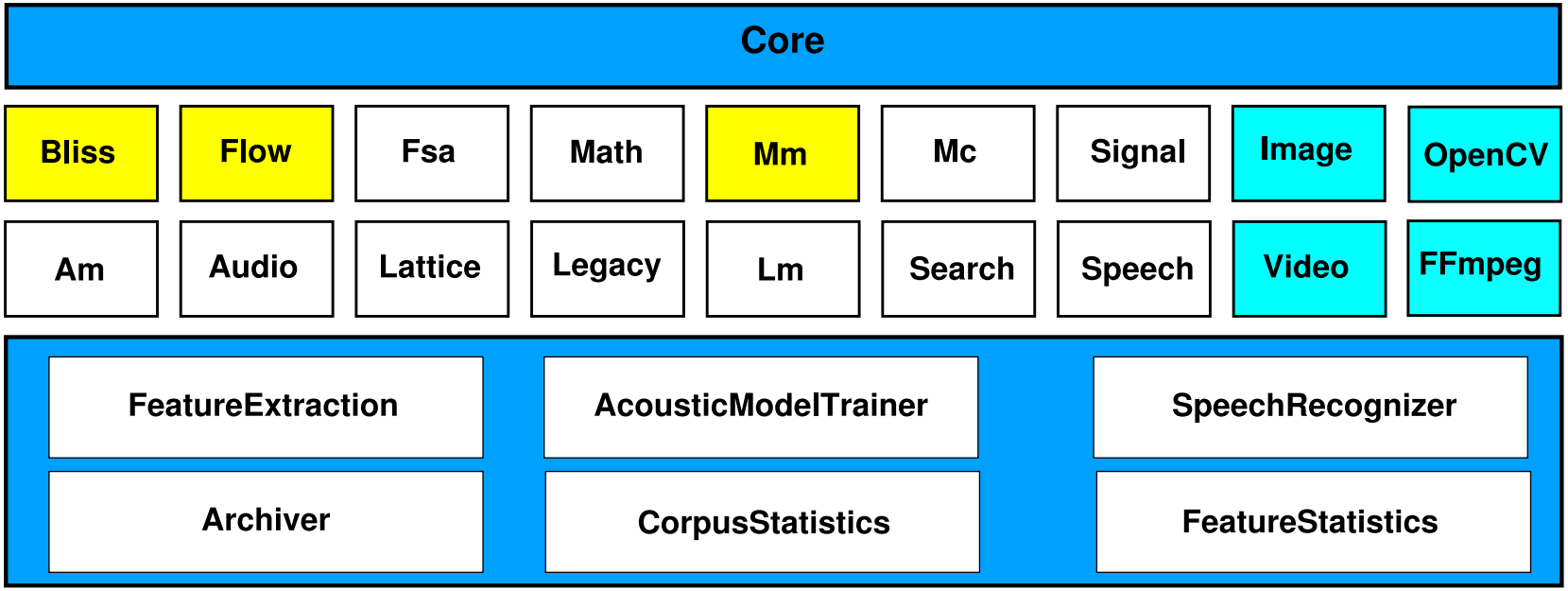


- corpus driven architecture, parallelization at segment-level
- runs on a 500-machine cluster (SUN Grid Engine)

RWTH OCR

Optical Character Recognition

Software



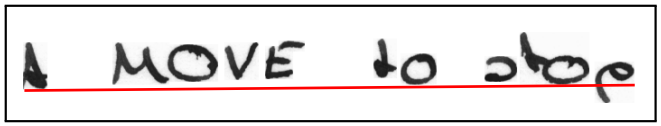
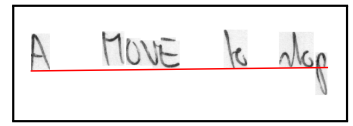
- corpus driven architecture, parallelization at segment-level
- runs on a 500-machine cluster (SUN Grid Engine)

$\Rightarrow$ `http://www.hltpr.rwth-aachen.de/rwth-ocr/`

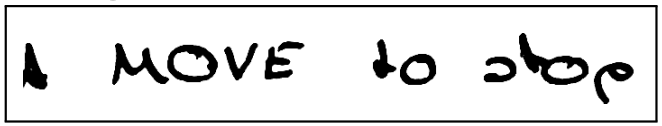
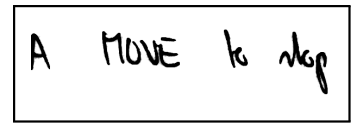
UPV Preprocessing - Latin

Optical Character Recognition

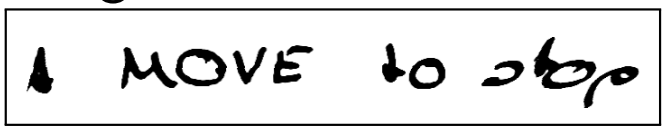
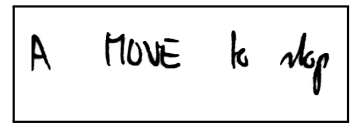
- Original images

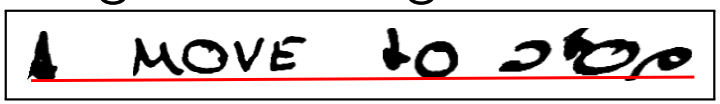
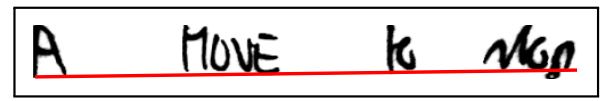
- Images after color normalisation

- Images after slant correction

- Images after height normalisation

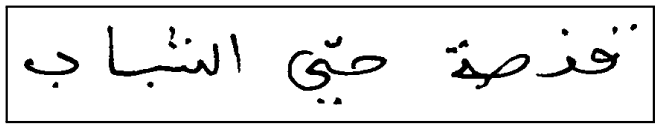
Note: preprocessing did not help for Arabic handwriting [Visualization]

◂ Return ▸ Features

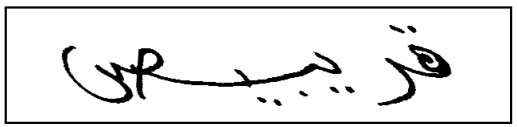
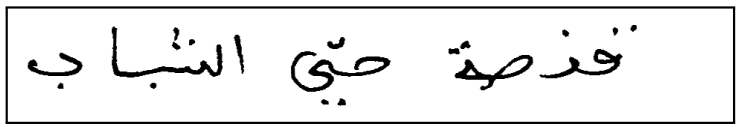
UPV Preprocessing - Arabic

Optical Character Recognition

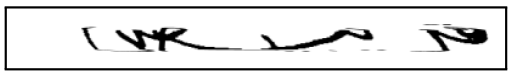
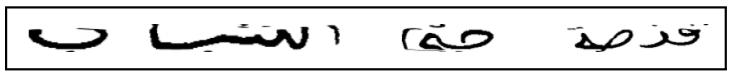
- Original images

- Images after slant correction

- Images after height normalisation

Experimental Results:

- important informations in ascender/descender areas lost

$\Rightarrow$ not yet suitable for Arabic OCR

◂ Return ▸ Features

RAW Features (values for IfN/ENIT)

- first level MLP system
  - input features: raw pixel column vectors ($\mathbf{32}$ components)
  - no windowing of input features
  - single hidden layer ($\mathbf{2000}$ nodes)
  - $\mathbf{216}$ output nodes (GDL glyph labels)
  - log-PCA transformation to $\mathbf{32}$ components
- second level MLP system
  - input features: concatenates MLP log-PCA with raw features
  - window size of $\mathbf{9} \Rightarrow (\mathbf{32} + \mathbf{32}) \times \mathbf{9} = \mathbf{576}$
  - single hidden layer ($\mathbf{3000}$ nodes)
  - $\mathbf{216}$ output nodes (GDL glyph labels)
  - log-PCA transformation to $\mathbf{32}$ components

◂ Return ▸ MLP Features ▸ MLP Modeling

TRAP-DCT Features (values for IfN/ENIT)

- first level MLP system
  - input features: raw pixel column vectors ($\mathbf{32}$ components)
  - TRAP-DCT [Hermansky & Sharma 98] window $\Rightarrow \mathbf{256}$ components
  - single hidden layer ($\mathbf{1500}$ nodes)
  - $\mathbf{216}$ output nodes (GDL glyph labels)
  - log-LDA transformation to $\mathbf{96}$ components
- second level MLP system
  - input features: MLP log-LDA with raw pixel features
  - two windows: $(\mathbf{96} \times \mathbf{5}) + (\mathbf{32} \times \mathbf{9}) = \mathbf{768}$
  - single hidden layer ($\mathbf{3000}$ nodes)
  - $\mathbf{216}$ output nodes (GDL glyph labels)
  - log-LDA transformation to $\mathbf{36}$ components

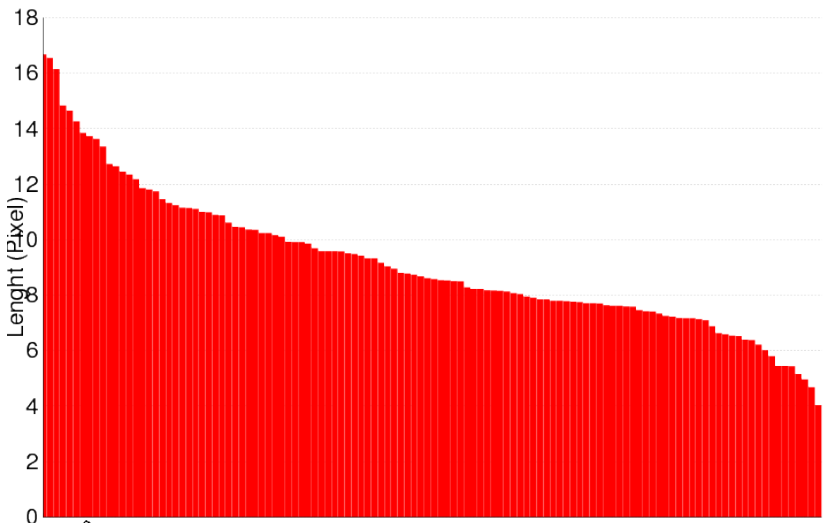
◂ Return ▸ MLP Features ▸ MLP Modeling

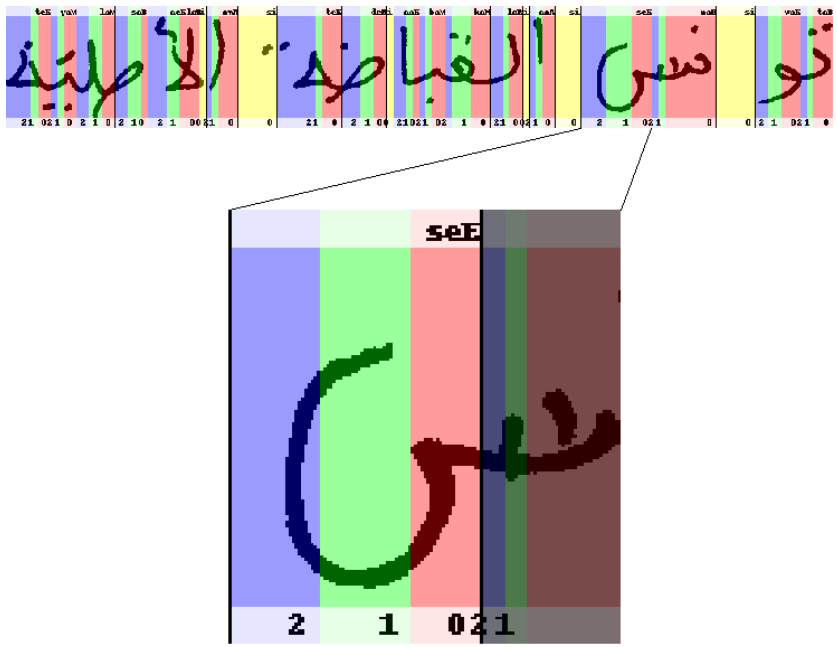
Glyph Dependent Lengths

Optical Character Recognition

Original Length

- overall mean of character length = 7.9 px (≈ 2.6 px/state)
- total #states = 357





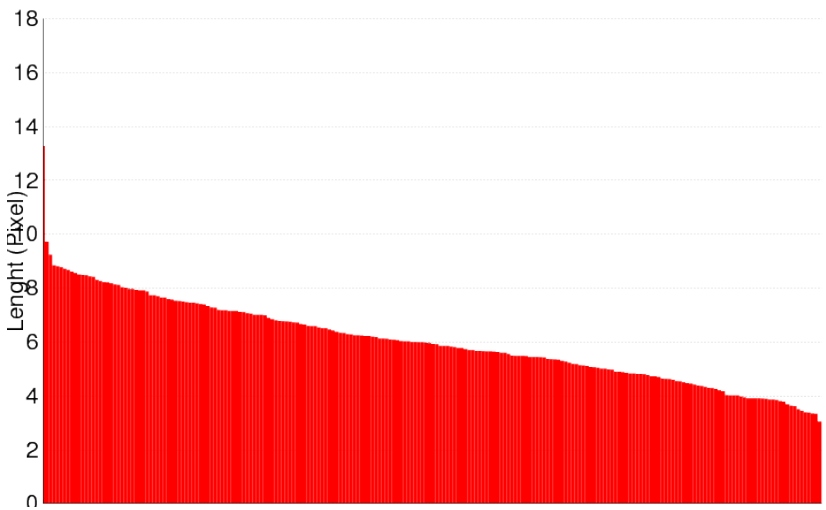
◂ Return ▸ GDL Modeling

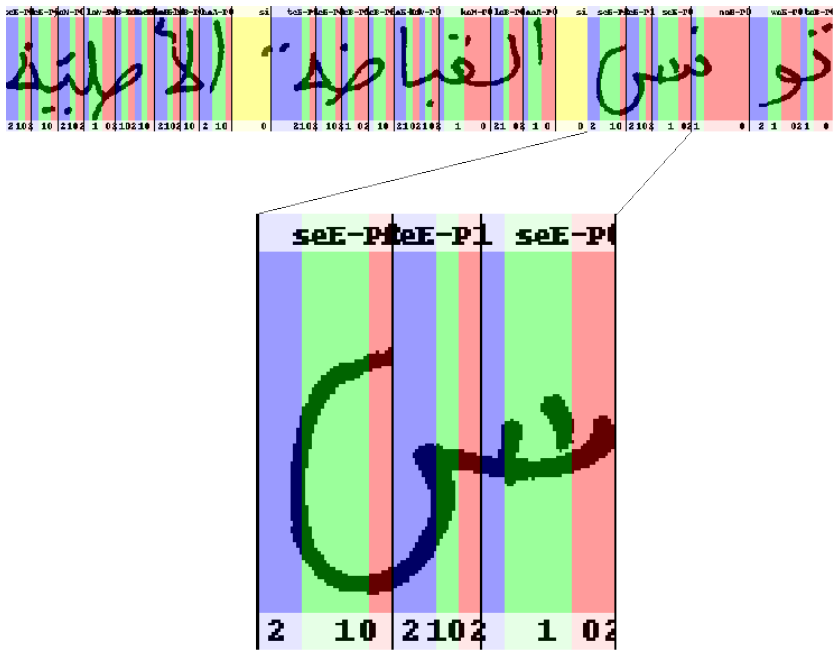
Glyph Dependent Lengths

Optical Character Recognition

Estimated Length

- overall mean of character length = 6.2 px (≈ 2.0 px/state)
- total #states = 558





◂ Return ▸ GDL Modeling

HMM baseline system

- searching for an unknown word sequence $w_1^N := w_1, \ldots, w_N$
- unknown number of words N
- maximize the posterior probability $p(w_1^N | x_1^T)$
- described by Bayes' decision rule:

$$x_1^T \rightarrow \hat{w}_1^N(x_1^T) = \arg\max_{w_1^N} \left\{ p^\kappa(w_1^N)\ p(x_1^T | w_1^N) \right\}$$

with κ a scaling exponent of the language model.

◂ Return ▸ Writing Variants Lexicon

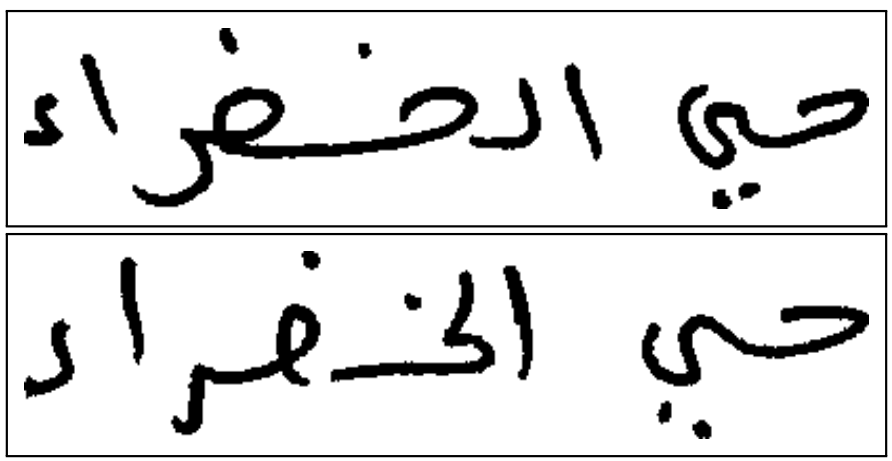
Writing Variant Model Refinement

Optical Character Recognition

Arabic Ligatures and Diacritics

- same Arabic word can be written in several writing variants
- depends on writer's handwriting style
  - $\Rightarrow$ lexicon with multiple writing variants
  - $\Rightarrow$ problem: many and rare writing variants

Example



◂ Return ▸ Writing Variants Lexicon

Writing Variant Model Refinement

Optical Character Recognition

- probability $p(v|w)$ for a variant v of a word w
  - usually considered as equally distributed
  - here: we use the count statistics as probability:

$$p(v|w) = \frac{\mathbf{N}(v,w)}{\mathbf{N}(w)}$$

- writing variant model refinement:

$$p(x_1^T|w_1^N) = \max_{v_1^N|w_1^N} \left\{ p^{\alpha}(v_1^N|w_1^N) p(x_1^T|v_1^N, w_1^N) \right\}$$

with v_1^N a sequence of unknown writing variants
α a scaling exponent of the writing variant probability

$\Rightarrow$ training: corpus and lexicon with supervised writing variants possible!

◂ Return ▸ Writing Variants Lexicon

Writer Adaptive Training

Optical Character Recognition

- writer adaptation
  - method for improving visual models in handwriting recognition
  - refine models by adaptation data of particular writers
  - widely used is affine transform based model adaptation

- CMLLR
  - Idea: normalize writing styles by adaptation of the features $\boldsymbol{x_t}$
  - constrained MLLR feature adaptation technique
  - also known as feature space MLLR (fMLLR) [Details]
  - estimate affine feature transform:

$$\boldsymbol{x}'_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{b}$$

  - CMLLR is text dependent
    - requires an (automatic) transcription

◂ Return ▸ Training / Decoding ▸ CMLLR

Writer Adaptive Training

Optical Character Recognition

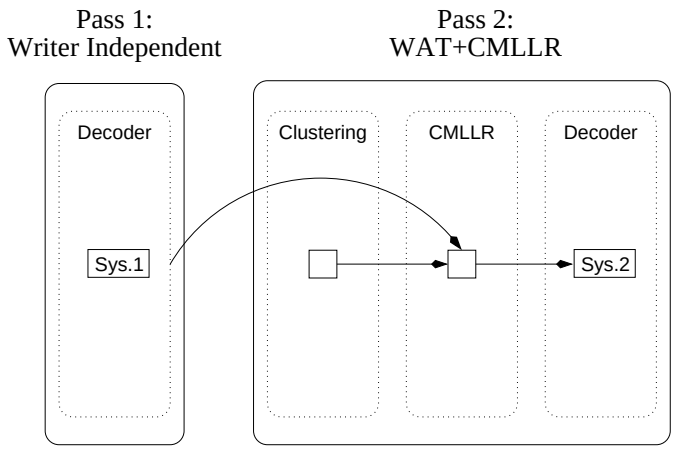
- writer adaptation compensates for writer differences during recognition
  - $\Rightarrow$ do the same during visual model training
  - $\Rightarrow$ maximize the performance gains from writer adaptation

- writer variations are compensated by writer adaptive training (WAT)
- writer normalization using CMLLR
- necessary steps
  1. train writer independent GMMs model
  2. CMLLR transformations are estimated for each (estimated) writer
     - supervised if writers are known
  3. apply CMLLR transformations on features to train writer dependent GMMs

Decoding: CMLLR-based Writer Adaptation

Optical Character Recognition

- writers and writing styles are unknown
- necessary steps
  1. estimate writing styles using clustering
     - Bayesian Information Criterion (BIC) based stopping condition
  2. estimate CMLLR feature transformations
     for every estimated writing style cluster
  3. second pass recognition
     - WAT models + CMLLR transformed features



Results - Decoding: Writer Adaptation

Optical Character Recognition

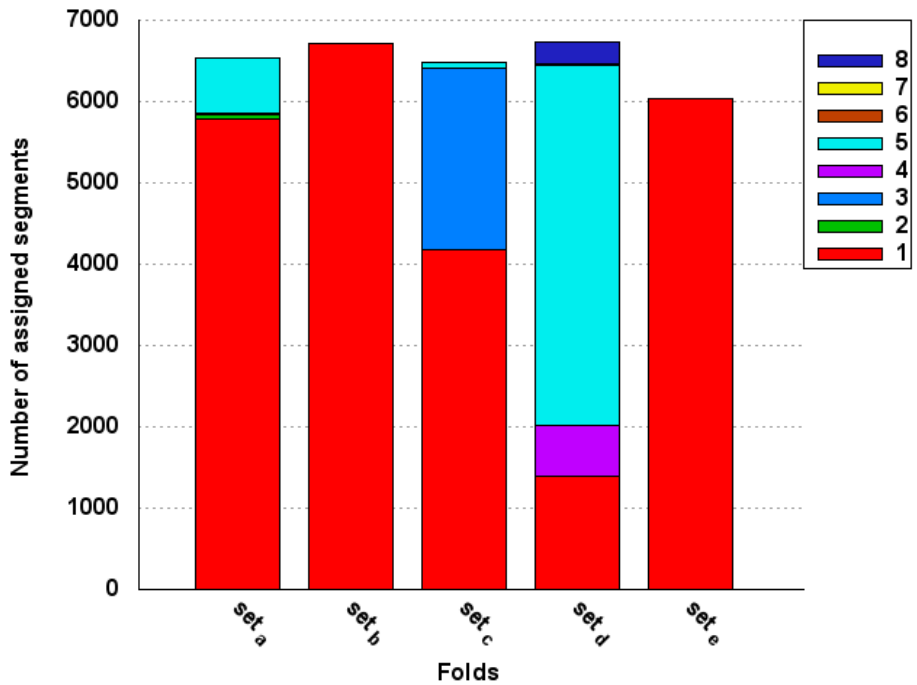
- comparison of GDL, WAT, and CMLLR based feature adaptation
- comparison of unsupervised and supervised writer clustering
  - decoding always unsupervised
  - supervised clustering $\Rightarrow$ only the writer labels are used!

Train	Test	WER[%]			
		1st pass		2nd pass	
		ML	+GDL	WAT+CMLLR	
				unsup.	sup.
abc	d	10.88	7.83	7.72	**5.82**
abd	c	11.50	8.83	9.05	**5.96**
acd	b	10.97	7.81	7.99	**6.04**
bcd	a	12.19	8.70	8.81	**6.49**
abcd	e	21.86	16.82	17.12	**11.22**

Results - Decoding: Writer Adaptation

Optical Character Recognition

- unsupervised clustering: error analysis
  - histograms for segment assignments over the different test folds
  - problem: unbalanced segment assignments



C-MLLR

Optical Character Recognition

Idea: improve the hypotheses by adaptation of the features $\boldsymbol{x_t}$

- effective algorithm for adaptation to a new speaker or environment (ASR)
- GMMs are used to estimate the CMLLR transform
- iterative optimization (ML criterion)
  - align each frame $\boldsymbol{x_t}$ to one HMM state (i.e. GMM)
  - accumulate to estimate the adaptation transform $\boldsymbol{A}$
  - likelihood function of the adaptation data given the model is to be maximized with respect to the transform parameters $\boldsymbol{A}, \boldsymbol{b}$
- one CMLLR transformation per (estimated) writer
- constrained refers to the use of the same matrix $\boldsymbol{A}$ for the transformation of the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$:

$$\boldsymbol{x'_t} = \boldsymbol{A x_t} + \boldsymbol{b} \Rightarrow \boldsymbol{N}(\boldsymbol{x}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \text{ with } \hat{\boldsymbol{\mu}} = \boldsymbol{A\mu} + \boldsymbol{b}$$

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{A\Sigma A^T}$$

◂ Return ▸ WAT

Bayes' Decision Rule and HMMs

$$x_1^T \rightarrow \hat{w}_1^N(x_1^T) = \arg\max_{w_1^N} \left\{ p^\kappa(w_1^N) \; p(x_1^T | w_1^N) \right\}$$

$$p(w_1^N) = \prod_{n=1}^{N} p(w_n | w_{n-1}^{n-m+1})$$

LMs

- any model in ARPA LM format can be read
- otherwise (weighted) finite state automatons
- typically:
  - modified Kneser-Ney smoothing
  - word LMs: 3- to 5-gram
  - character LMs: 5- to 10-gram

Discriminative Training

Optical Character Recognition

Goals for OCR

- can we adopt from ASR? parameter behavior?
- novel: unsupervised adaptation possible?

$\Rightarrow$ joint work with Georg Heigold, details in his PhD Thesis [Heigold 10]

Introduction

- labeled training sentences $(\boldsymbol{X_r}, \boldsymbol{W_r})_{r=1,\ldots,R}$ with
  2D image $\Rightarrow$ string representation $\boldsymbol{X} = \boldsymbol{x_1}, \ldots, \boldsymbol{x_T}$
  word sequence $\boldsymbol{W} = \boldsymbol{w_1}, \ldots, \boldsymbol{w_N}$
  $\boldsymbol{p_\Lambda}(\boldsymbol{X}, \boldsymbol{W})$ with model parameters $\boldsymbol{\Lambda} \Rightarrow$ posterior

$$p_{\Lambda,\gamma}(W|X) = \frac{p_\Lambda(X,W)^\gamma}{\sum_V p_\Lambda(X,V)^\gamma}$$

Introduction

- labeled training sentences $(\boldsymbol{X}_r, \boldsymbol{W}_r)_{r=1,\ldots,R}$
- training: weighted accumulation of aligned observations $\boldsymbol{x}_t$:

$$\text{accumulator}_s = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \omega_{r,s,t} \cdot \boldsymbol{x}_t$$

Introduction

- labeled training sentences $(X_r, W_r)_{r=1,\ldots,R}$
- training: weighted accumulation of aligned observations x_t:

$$\text{accumulator}_s = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \omega_{r,s,t} \cdot x_t$$

Maximum Mutual Information (MMI)

$$\omega_{r,s,t} := \frac{p(X_r, W_r)^\gamma}{\sum_V p(X_r, V)^\gamma}$$

- $\omega_{r,s,t}$ is the "(true) posterior" weight

Discriminative Training

Optical Character Recognition

Margin-Based MMI (M-MMI)

$$\omega_{r,s,t}(\rho \neq 0) := \frac{\left\{p(X_r, W_r)\, e^{-\rho A(W_r, W_r)}\right\}^{\gamma}}{\sum_{V} \left\{p(X_r, W_r)\, e^{-\rho A(V, W_r)}\right\}^{\gamma}}$$

- additional margin-term including the accuracy $A(\cdot, W_r)$
  e.g. approximate word error [Povey & Woodland 02]
- $\omega_{r,s,t}$ is the "margin posterior" weight

[Heigold & Dreuw+ 10], IEEE J-STSP

Discriminative Training

Optical Character Recognition

Unsupervised Discriminative Model Adaptation

- assumption: margin-based training robust against outliers

$\Rightarrow$ unsupervised discriminative training on test data

$\Rightarrow$ select data depending on confidence threshold τ_c

[Dreuw & Heigold+ 11], IJDAR

Confidence-Based Accumulation

1. recognize (unsupervised transcriptions)
2. estimate frame confidences $c_{r,s,t}$ (state-posteriors, FB algo.)
3. 1-best accumulation: consider only observations for which $c_{r,s,t} > \tau_c$ in the 1-best recognition hypothesis

[Gollan & Bacchiani 08]

Confidence-Based M-MMI (M-MMI-conf)

- sentence/word confidences $\Rightarrow$ simply weight the segments
- state confidences $\Rightarrow$ state posteriors required

$$\omega_{r,s,t} := \frac{\left\{ \sum_{s_1^{T_r}: s_t = s} p(X_r, s_1^{T_r}, W_r) \cdot \exp(-\rho A(W_r, W_r)) \right\}^{\gamma}}{\underbrace{\sum_{V} \left\{ \sum_{s_1^{T_r}: s_t = s} p(X_r, s_1^{T_r}, V) \cdot \underbrace{\exp(-\rho A(V, W_r))}_{\text{margin}} \right\}^{\gamma}}_{\text{posterior}}} \cdot \underbrace{\delta(c_{r,s,t} > \tau_c)}_{\text{confidence}}$$

$\Rightarrow$ accumulator$_s$:
each frame t contributes $\omega_{r,s,t} \cdot \delta(c_{r,s,t} > \tau_c) \cdot x_t$

Training Criteria

Optical Character Recognition

Maximum Likelihood: accumulation of aligned $\boldsymbol{x_t}$

$$\text{accumulator}_s = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \delta(s_t, s) \cdot \boldsymbol{x_t}$$

Margin-based MMI/MPE: weighted accumulation

$$\text{accumulator}_s = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \omega_{r,s,t}(\rho) \cdot \boldsymbol{x_t}$$

Confidence-Based M-MMI/M-MPE: confidence-weighted accumulation

$$\text{accumulator}_s = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \omega_{r,s,t}(\rho) \cdot \delta(c_{r,s,t} > \tau_c) \cdot \boldsymbol{x_t}$$

- with $c_{r,s,t}$ at sentence-, word-, glyph-, or state-level

Optimization Problem - Loss Minimization

- loss function for each training sample r:

$$L[p_\Lambda(X_r, \cdot), W_r]$$

- criterion

$$\hat{\Lambda} = \arg\min_\Lambda \left\{ C||\Lambda - \Lambda_0||_2^2 + \sum_{r=1}^{R} L[p_\Lambda(X_r, \cdot), W_r] \right\}$$

  - ℓ_2 regularization term replaced by I-smoothing [Povey & Woodland 02]
  - initialization at a reasonable ML trained model Λ_0

◂ Return ▸ Discriminative Training

Discriminative Training

Optical Character Recognition

Maximum Mutual Information (MMI)

$$L^{(\text{MMI})}[p_\Lambda(X_r, \cdot), W_r] =$$

$$- \log \frac{p_\Lambda(X_r, W_r)^\gamma}{\sum_V p_\Lambda(X_r, V)^\gamma}$$

Discriminative Training

Optical Character Recognition

Maximum Mutual Information (MMI)

$$L^{(\text{MMI})}[p_\Lambda(X_r, \cdot), W_r] = - \log \frac{p_\Lambda(X_r, W_r)^\gamma}{\sum_V p_\Lambda(X_r, V)^\gamma}$$

Margin-Based MMI (M-MMI)

$$L_\rho^{(\text{M-MMI})}[p_\Lambda(X_r, \cdot), W_r] = - \log \frac{[p_\Lambda(X_r, W_r)\ \exp(-\rho A(W_r, W_r))]^\gamma}{\sum_V [p_\Lambda(X_r, V)\ \exp(-\rho A(V, W_r))]^\gamma}$$

- additional margin-term including the accuracy $A(\cdot, W_r)$ e.g. approximate word error [Povey & Woodland 02]

Minimum Phone Error (MPE)

$$L^{(\text{MPE})}[p_\Lambda(X_r, \cdot), W_r] =$$

$$\sum_W E(W, W_r) \frac{p_\Lambda(X_r, W_r)^\gamma}{\sum_V p_\Lambda(X_r, V)^\gamma}$$

- error function $E(\cdot, W_r)$, e.g. approximate phone error [Povey & Woodland 02]

Discriminative Training

Optical Character Recognition

Minimum Phone Error (MPE)

$$L^{(\text{MPE})}[p_\Lambda(X_r, \cdot), W_r] =$$

$$\sum_W E(W, W_r) \frac{p_\Lambda(X_r, W_r)^\gamma}{\sum_V p_\Lambda(X_r, V)^\gamma}$$

- error function $E(\cdot, W_r)$, e.g. approximate phone error [Povey & Woodland 02]

Margin-Based MPE (M-MPE)

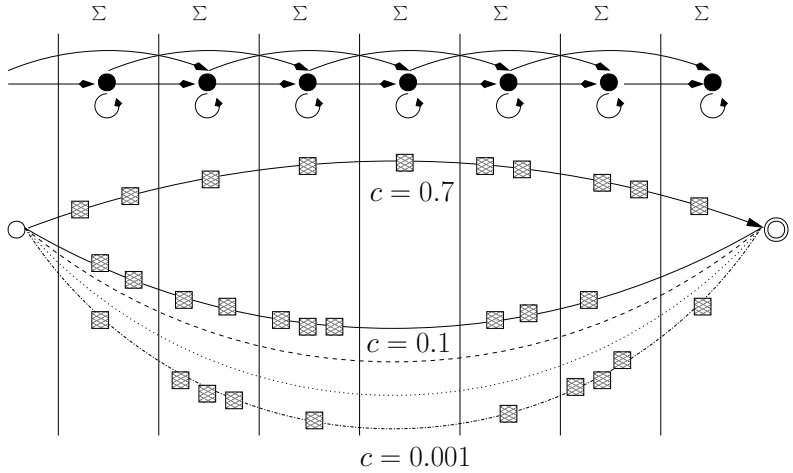
$$L_\rho^{(\text{M-MPE})}[p_\Lambda(X_r, \cdot), W_r] =$$

$$\sum_W E(W, W_r) \frac{[p_\Lambda(X_r, W_r) \ \exp(-\rho A(W, W_r))]^\gamma}{\sum_V [p_\Lambda(X_r, V) \ \exp(-\rho A(V, W_r))]^\gamma},$$

M-MMI-conf Training

Optical Character Recognition

- example for a word-graph w/ 1-best state alignment



- steps for confidence-based model adaptation:
  - 1-pass recognition (unsupervised transcriptions)
  - calculation of corresponding confidences
  - unsupervised M-MMI-conf training on test data to adapt models (w/ regularization)
- can be done iteratively with unsupervised corpus update!

◂ Return ▸ Training and Decoding

Discriminative Training

Optical Character Recognition

Confidence-Based M-MMI (M-MMI-conf)

- sentence/word confidences $\Rightarrow$ simply weight the segments
- state confidences $\Rightarrow$ state posteriors required

$$\omega_{r,s,t} := \frac{\underbrace{\left\{ \sum_{s_1^{T_r}: s_t = s} p(X_r, s_1^{T_r}, W_r) \cdot \exp(-\rho A(W_r, W_r)) \right\}^{\gamma}}_{\text{posterior}}}{\sum_V \underbrace{\left\{ \sum_{s_1^{T_r}: s_t = s} p(X_r, s_1^{T_r}, V) \cdot \underbrace{\exp(-\rho A(V, W_r))}_{\text{margin}} \right\}^{\gamma}}_{\text{posterior}}} \cdot \underbrace{\delta(c_{r,s,t} > \tau_c)}_{\text{confidence}}$$

$\Rightarrow$ accumulator acc_s:
each frame t contributes $\omega_{r,s,t} \cdot \delta(c_{r,s,t} > \tau_c) \cdot x_t$

Approximate Phone Error

- proposed by Povey [Povey & Woodland 02]
- phone accuracy of a word sequence $\boldsymbol{W}$

$\Rightarrow$ sum over all phone arcs $\boldsymbol{q}$ in the sequence $\boldsymbol{W}$

$$\text{PhoneAcc}(\boldsymbol{q}|\boldsymbol{W}) = \max_{\boldsymbol{z} \in \boldsymbol{W}} \begin{cases} -\mathbf{1} + \mathbf{2}\boldsymbol{e}(\boldsymbol{q}|\boldsymbol{z}), & \text{if same phone} \\ -\mathbf{1} + \boldsymbol{e}(\boldsymbol{q}|\boldsymbol{z}), & \text{if different} \end{cases}$$

- $\boldsymbol{q}$ hyp, $\boldsymbol{z}$ reference, $\boldsymbol{e}$ overlap in time

$\Rightarrow$ efficiently calculated by pre-computing for each frame a list of arcs that include that frame

$\Rightarrow$ approximate word error similar (e.g. for M-MMI or MWE)

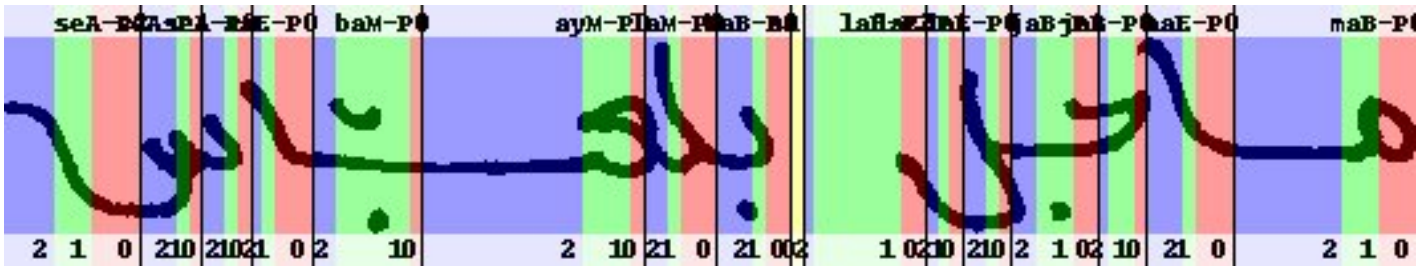
◂ Return ▸ MMI/M-MMI ▸ MPE/M-MPE

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Visual Inspections

- ML

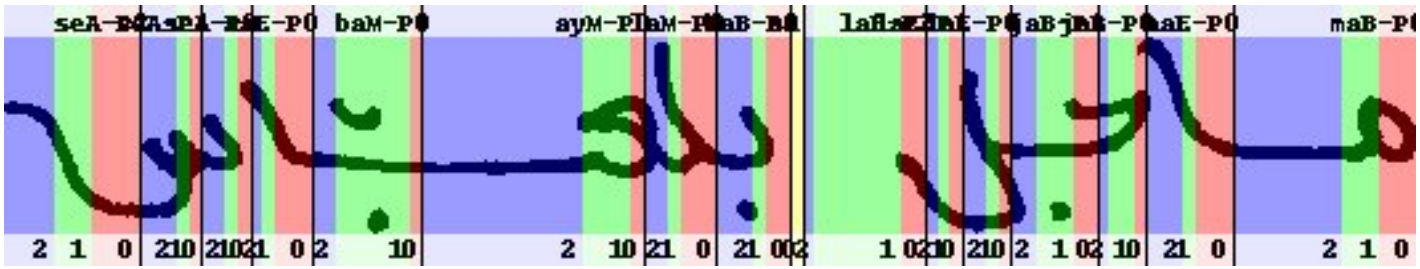


Isolated Arabic Words - IfN/ENIT

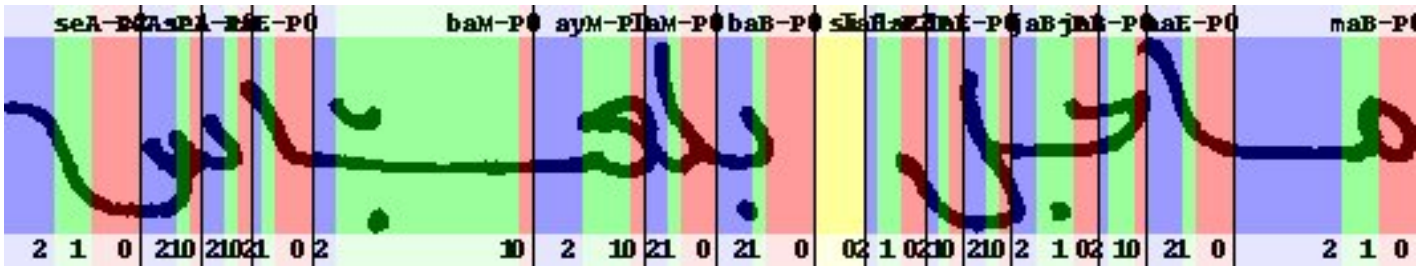
Optical Character Recognition

Visual Inspections

- ML



- M-MMI-conf



⇒ learned to discriminate depending on white-space context

⇒ implicit HMM segmentation adequate for post-processing

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

- 937 Tunisian city names
- 32492 handwritten Arabic words, about 1000 writers
- database is used by more than 60 groups all over the world

- writer statistics

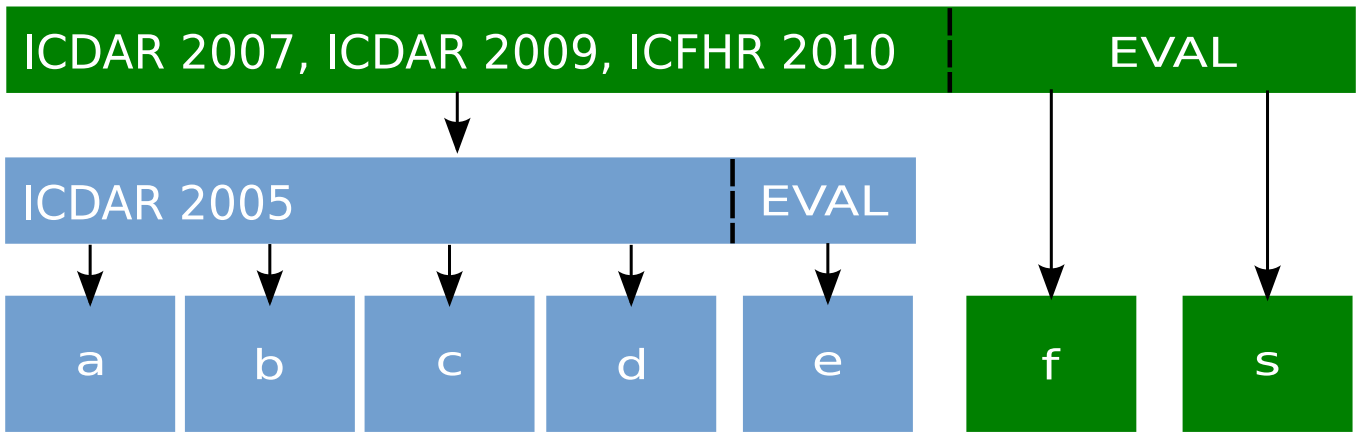
set	#writers	#samples
a	0.1k	6.5k
b	0.1k	6.7k
c	0.1k	6.4k
d	0.1k	6.7k
e	0.5k	6.0k
f	-	8.6k
s	-	1.5k

- examples (same word):

Return

Competitions and Corpus Development

- external evaluations
- ICDAR 2005: a-d sets for training, evaluation on set e
- ICDAR 2007: a-e sets for training, evaluation on set f, s
  - set f from same Tunisian University
  - set s from United Arab Emirates
- ICDAR 2009 and ICFHR 2010: as for ICDAR 2007



Return

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Baseline System

- appearance-based sliding window features + PCA
- ML trained GHMM: 121 glyphs, 361 GMMs, 36k densities

Train	Test	WER[%]				
		1st pass				2nd pass
		ML	GDL	+MMI	+M-MMI	M-MMI-conf
abc	d	10.9				
abd	c	11.5				
acd	b	11.0				
bcd	a	12.2				
abcd	e	21.9				

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Baseline System

- appearance-based sliding window features + PCA
- ML trained GHMM: 121 glyphs, 361 GMMs, 36k densities
- + GDL: 216 glyphs, 646 GMMs, 55k densities

Train	Test	WER[%]				
		1st pass				2nd pass
		ML	GDL	+MMI	+M-MMI	M-MMI-conf
abc	d	10.9	**7.8**			
abd	c	11.5	**8.8**			
acd	b	11.0	**7.8**			
bcd	a	12.2	**8.7**			
abcd	e	21.9	**16.8**			

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Baseline System

- appearance-based sliding window features + PCA
- ML trained GHMM: 121 glyphs, 361 GMMs, 36k densities
- + GDL: 216 glyphs, 646 GMMs, 55k densities

Train	Test	WER[%]				
		1st pass				2nd pass
		ML	GDL	+MMI	+M-MMI	M-MMI-conf
abc	d	10.9	7.8	7.4	**6.1**	
abd	c	11.5	8.8	8.2	**6.8**	
acd	b	11.0	7.8	7.6	**6.1**	
bcd	a	12.2	8.7	8.4	**7.0**	
abcd	e	21.9	16.8	16.4	**15.4**	

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Baseline System

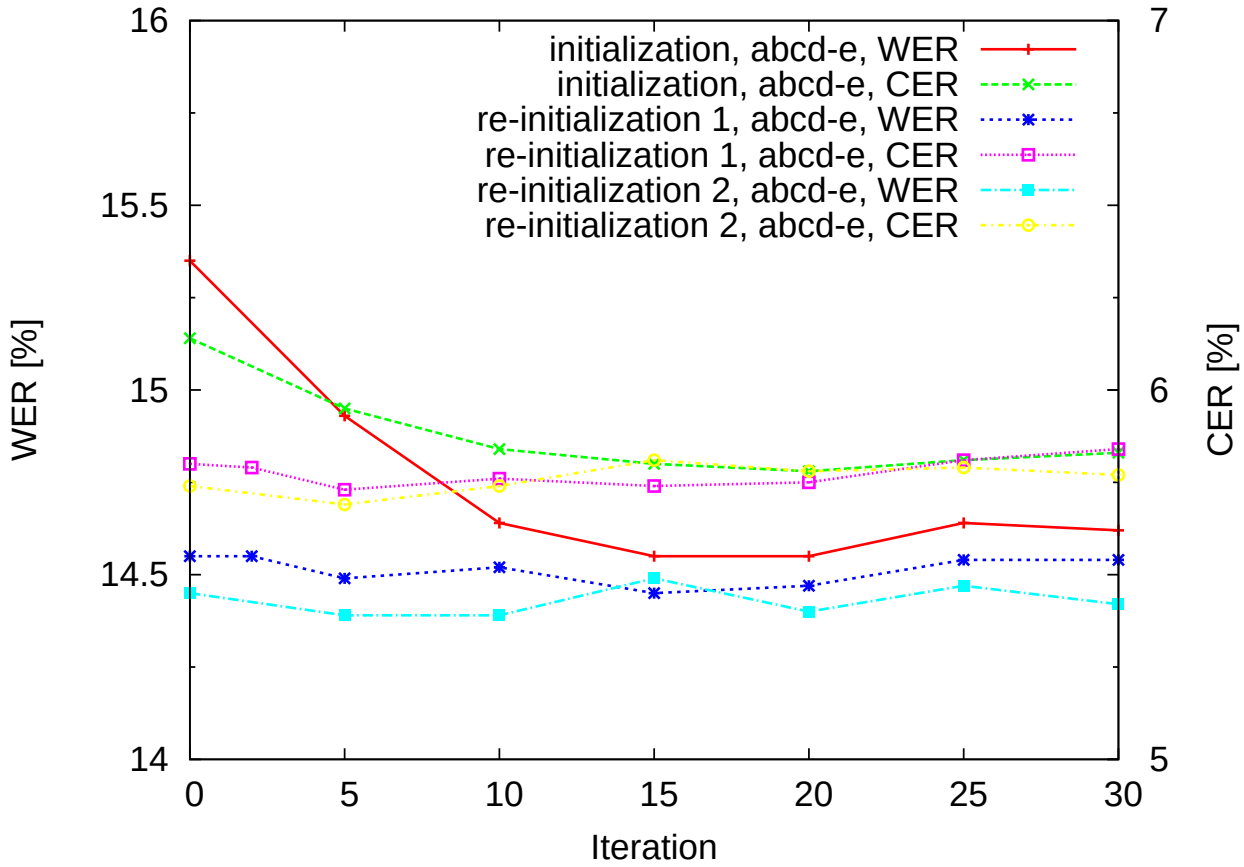
- appearance-based sliding window features + PCA
- ML trained GHMM: 121 glyphs, 361 GMMs, 36k densities
- + GDL: 216 glyphs, 646 GMMs, 55k densities

Train	Test	WER[%]				
		1st pass				2nd pass
		ML	GDL	+MMI	+M-MMI	M-MMI-conf
abc	d	10.9	7.8	7.4	6.1	**6.0**
abd	c	11.5	8.8	8.2	6.8	**6.4**
acd	b	11.0	7.8	7.6	6.1	**5.8**
bcd	a	12.2	8.7	8.4	7.0	**6.8**
abcd	e	21.9	16.8	16.4	15.4	**14.6**

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Unsupervised Training



Hybrid MLP/HMM vs. Tandem MLP-GHMM

- both GHMM systems are M-MMI trained

Model	WER[%]	CER[%]
GHMM	15.4	6.1
MLP/HMM	11.6	4.8
MLP-GHMM	**7.3**	**3.0**

$\Rightarrow$ MLP based features (hybrid/tandem) very powerful!

$\Rightarrow$ tandem usually outperforms hybrid approaches

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Hybrid MLP/HMM vs. Tandem MLP-GHMM

Train	Test	GHMM		MLP/HMM		MLP-GHMM	
		WER[%]	CER[%]	WER[%]	CER[%]	WER[%]	CER[%]
abc	d	6.1	2.4				
abd	c	6.8	2.6				
acd	b	6.1	2.2				
bcd	a	7.0	3.1				
abcd	e	15.4	6.1				

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Hybrid MLP/HMM vs. Tandem MLP-GHMM

- MLP parameters tuned only on set abc

Train	Test	GHMM		MLP/HMM		MLP-GHMM	
		WER[%]	CER[%]	WER[%]	CER[%]	WER[%]	CER[%]
abc	d	6.1	2.4	**4.5**	**1.7**		
abd	c	6.8	2.6	**2.6**	**0.9**		
acd	b	6.1	2.2	**2.7**	**0.9**		
bcd	a	7.0	3.1	**3.1**	**1.3**		
abcd	e	15.4	6.1	**11.6**	**4.5**		

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Hybrid MLP/HMM vs. Tandem MLP-GHMM

- MLP parameters tuned only on set abc
- both GHMM systems are M-MMI trained

Train	Test	GHMM		MLP/HMM		MLP-GHMM	
		WER[%]	CER[%]	WER[%]	CER[%]	WER[%]	CER[%]
abc	d	6.1	2.4	4.5	1.7	**3.5**	**1.5**
abd	c	6.8	2.6	2.6	0.9	**1.4**	**0.8**
acd	b	6.1	2.2	2.7	**0.9**	**2.5**	1.0
bcd	a	7.0	3.1	3.1	1.3	**2.6**	**1.1**
abcd	e	15.4	6.1	11.6	4.5	**7.3**	**3.0**

$\Rightarrow$ MLP based features (hybrid/tandem) very powerful!

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Comparisons: ICDAR / ICFHR Competitions

external evaluations on unknown sets f and s [Märgner & Abed 10]

Year	Group (Approach)	set-e	set-f	set-s
ICDAR 2007	SIEMENS (HMM)	18.1	12.8	26.1
	MIE (DP)	-	16.7	31.6
	UOB-ENST (HMM)	-	18.1	30.1

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Comparisons: ICDAR / ICFHR Competitions

external evaluations on unknown sets f and s [Märgner & Abed 10]

Year	Group (Approach)	set-e	set-f	set-s
ICDAR 2007	SIEMENS (HMM)	18.1	12.8	26.1
	MIE (DP)	-	16.7	31.6
	UOB-ENST (HMM)	-	18.1	30.1
ICDAR 2009	MDLSTM (RNN/CTC)	-	**6.6**	18.9
	A2iA (combined)	-	10.6	23.3
	RWTH OCR (HMM, M-MMI)	15.4	14.5	28.7
	RWTH OCR (HMM, M-MMI-conf)	14.6	14.3	27.5
	UOB-ENST (HMM, combined)	-	16.0	27.7

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Comparisons: ICDAR / ICFHR Competitions

external evaluations on unknown sets f and s [Märgner & Abed 10]

Year	Group (Approach)	set-e	set-f	set-s
ICDAR 2007	SIEMENS (HMM)	18.1	12.8	26.1
	MIE (DP)	-	16.7	31.6
	UOB-ENST (HMM)	-	18.1	30.1
ICDAR 2009	MDLSTM (RNN/CTC)	-	**6.6**	18.9
	A2iA (combined)	-	10.6	23.3
	(HMM)	-	17.8	33.6
	(MLP/HMM)	-	14.4	29.6
	RWTH OCR (HMM, M-MMI)	15.4	14.5	28.7
	RWTH OCR (HMM, M-MMI-conf)	14.6	14.3	27.5
	UOB-ENST (HMM, combined)	-	16.0	27.7

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Comparisons: ICDAR / ICFHR Competitions

external evaluations on unknown sets f and s [Märgner & Abed 11]

Year	Group (Approach)	set-e	set-f	set-s
ICFHR 2010	UPV PRHLT (HMM)	6.2	7.8	**15.4**
	RWTH OCR (x16, MLP-GHMM, M-MMI)	7.3	9.1	18.9
	UPV PRHLT (HMM, w/o vert. norm.)	12.3	12.1	21.6
	CUBS-AMA (HMM)	-	19.7	32.1

Isolated Arabic Words - IfN/ENIT

Optical Character Recognition

Comparisons: ICDAR / ICFHR Competitions

external evaluations on unknown sets f and s [Märgner & Abed 11]

Year	Group (Approach)	set-e	set-f	set-s
ICFHR 2010	UPV PRHLT (HMM)	6.2	7.8	**15.4**
	RWTH OCR (x16, MLP-GHMM, M-MMI)	7.3	9.1	18.9
	UPV PRHLT (HMM, w/o vert. norm.)	12.3	12.1	21.6
	CUBS-AMA (HMM)	-	19.7	32.1
ICDAR 2011	RWTH OCR (x32, MLP-GHMM, ML)	**5.9**	7.8	15.5
	REGIM (HMM)	-	21.0	31.6
	JU-OCR (RF & Rules)	-	36.1	50.3
	CENPARMI (SVMs)	-	60.0	64.5

$\Rightarrow$ missing is an M-MMI trained MLP-GHMM system!

Continuous Latin Lines - IAM

Optical Character Recognition

- English handwriting
- LM: Brown, Lancester-Oslo-Bergen, and Wellington corpora
- 50k lexicon, 3-gram LM

	Train	Devel	Eval	LM
words	53.8k	8.7k	25.4k	3.3M
chars	219.7k	31.7k	96.6k	13.8M
lines	6.1k	0.9k	2.7k	164k
writers	283	57	162	-
OOV rate	1.07%	3.94%	3.42%	1.87%

Continuous Latin Lines - IAM

Optical Character Recognition

Return

TODO: PPL plot

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Results

Systems	WER [%]		CER [%]	
	Devel	Eval	Devel	Eval
GHMM, ML baseline [Jonas 09]	31.9	38.9	8.4	11.8
+ M-MMI	25.8	31.6	7.6	11.8
+M-MMI-conf	23.7	29.0	6.8	10.5
+ M-MPE	24.3	30.0	6.9	10.9
+ M-MPE-conf	23.7	29.2	6.5	10.3
MLP/HMM	31.2	36.9	10.0	14.2
MLP-GHMM	25.7	32.9	7.7	12.4
+ M-MMI	23.5	30.1	6.7	11.1
+ M-MPE	22.7	28.8	6.1	10.1
[Bertolami & Bunke 08a] (GHMMs)	26.8	32.8	-	-
[Graves & Liwicki^{+} 09] (LSTM/CTC)	-	25.9	-	18.2
[Espana-Boquera & Castro-Bleda^{+} 11] (MLPs/HMM)	19.0	22.4	-	9.8

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

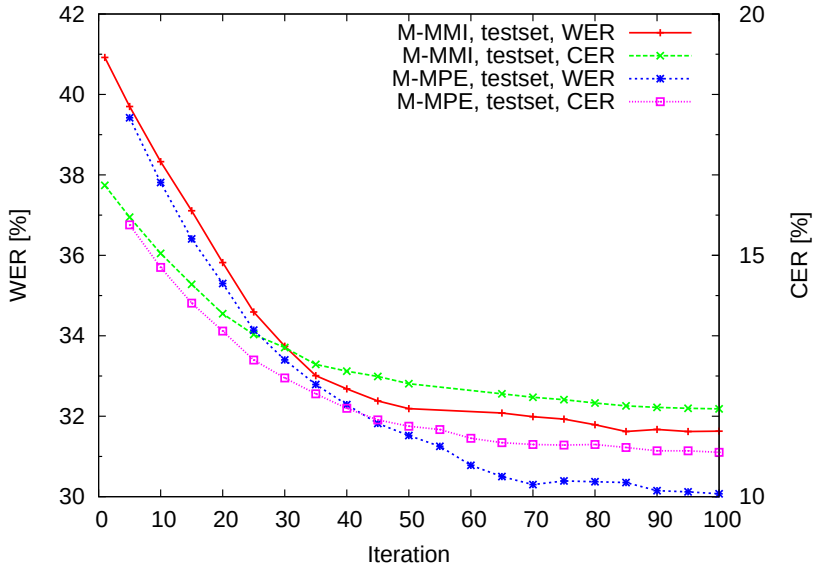
Results

Systems	WER [%]		CER [%]	
	Devel	Eval	Devel	Eval
GHMM, ML baseline [Jonas 09]	31.9	38.9	8.4	11.8
+ M-MMI	25.8	31.6	7.6	11.8
+M-MMI-conf	23.7	29.0	6.8	10.5
+ M-MPE	24.3	30.0	6.9	10.9
+ M-MPE-conf	23.7	29.2	6.5	10.3
MLP/HMM	31.2	36.9	10.0	14.2
MLP-GHMM	25.7	32.9	7.7	12.4
+ M-MMI	23.5	30.1	6.7	11.1
+ M-MPE	22.7	28.8	6.1	10.1
[Bertolami & Bunke 08a] (GHMMs)	26.8	32.8	-	-
[Graves & Liwicki^{+} 09] (LSTM/CTC)	-	25.9	-	18.2
[Espana-Boquera & Castro-Bleda^{+} 11] (MLPs/HMM)	19.0	22.4	-	9.8
[Doetsch 11] RWTH OCR (LSTM-GHMM, M-MPE)	**17.4**	**21.4**	**6.6**	**9.5**

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Margin-Based Supervised Training



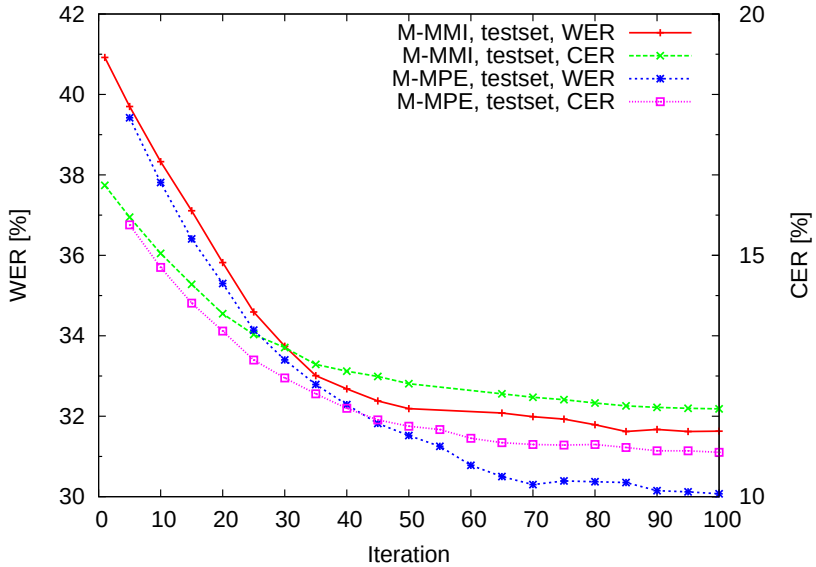
$\Rightarrow$ M-MPE usually outperforms M-MMI

$\Rightarrow$ benefit of margin term is significant (limited in ASR)

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Margin-Based Supervised Training



⇒ M-MPE usually outperforms M-MMI

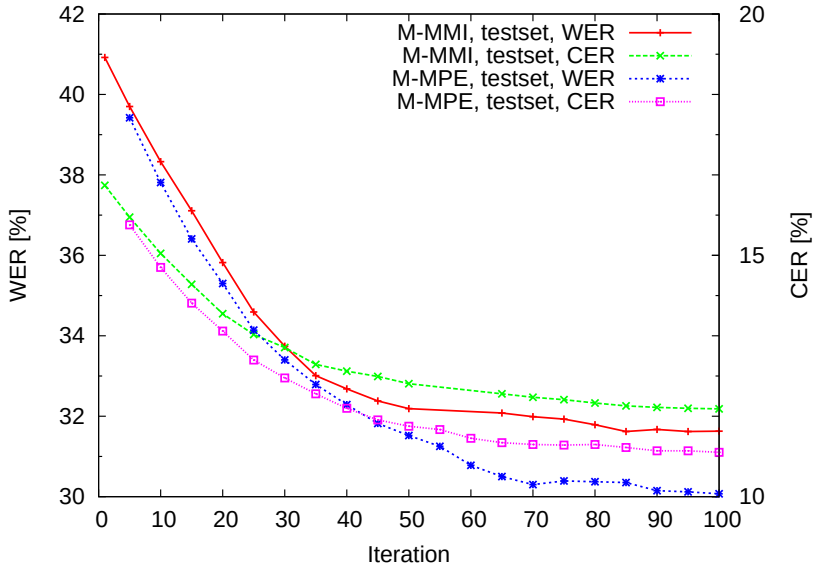
⇒ benefit of margin term is significant (limited in ASR)

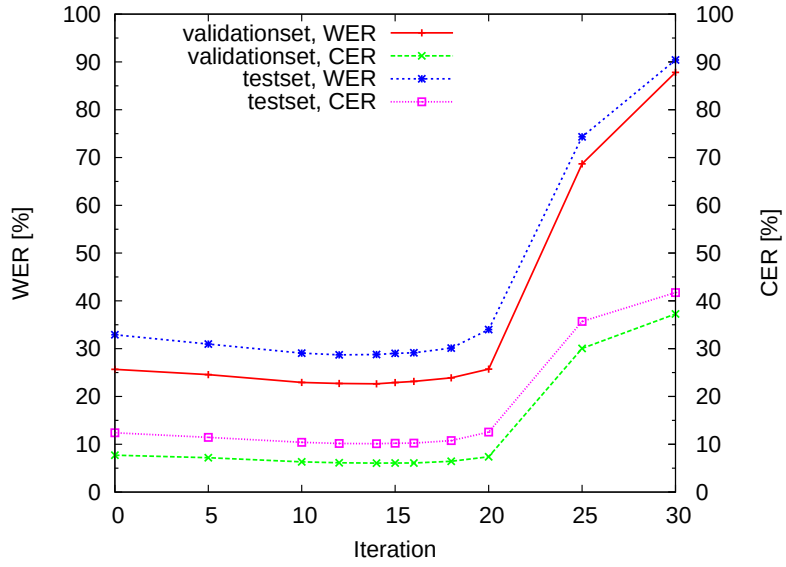
⇒ M-MPE word lattice density is important for convergence!

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Margin-Based Supervised Training





⇒ M-MPE usually outperforms M-MMI

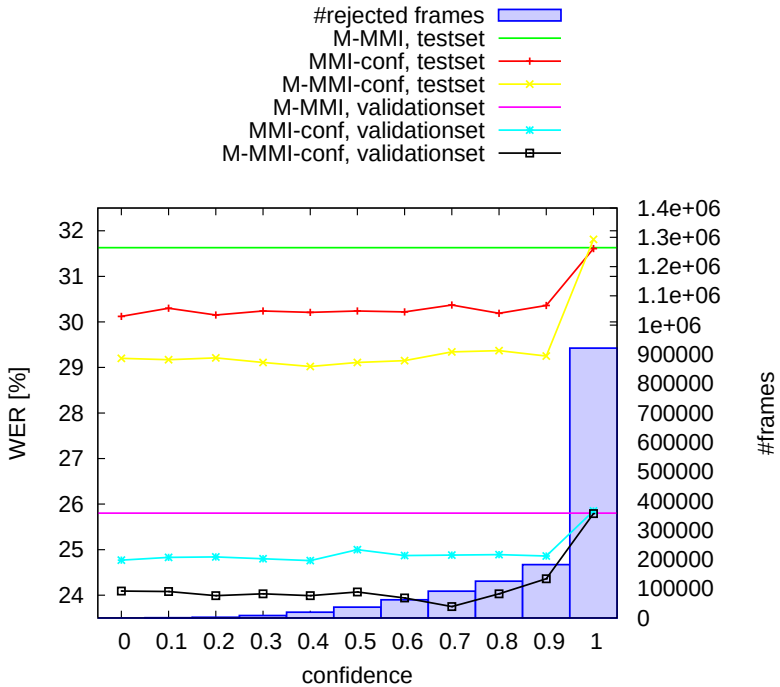
⇒ benefit of margin term is significant (limited in ASR)

⇒ M-MPE word lattice density is important for convergence!

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Margin- and Confidence-Based Unsupervised Training



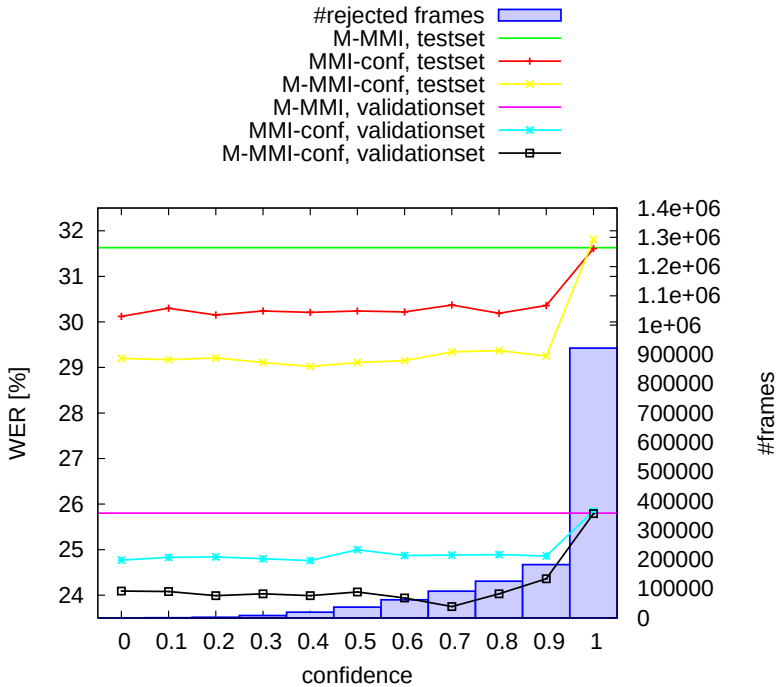
⇒ benefit of confidence term is limited, margin again significant

⇒ typically M-MMI/M-MMI-conf more robust

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Margin- and Confidence-Based Unsupervised Training



⇒ benefit of confidence term is limited, margin again significant

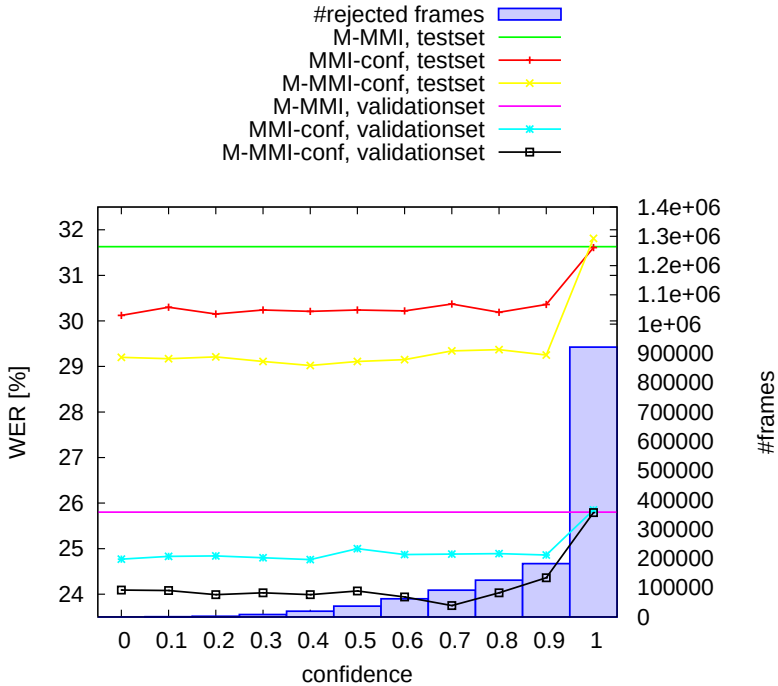
⇒ typically M-MMI/M-MMI-conf more robust

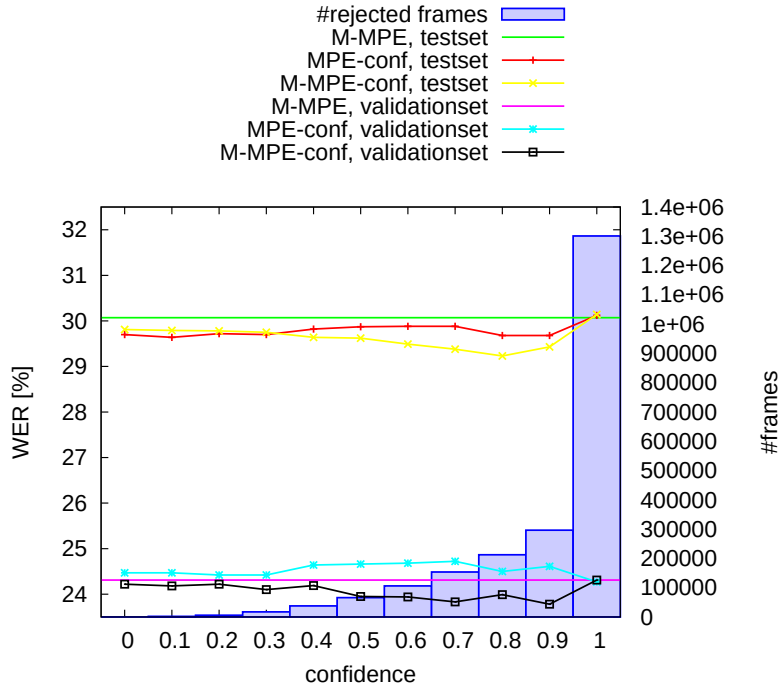
⇒ confidence term is important in M-MPE-conf

Continuous Latin Sentence Lines - IAM

Optical Character Recognition

Margin- and Confidence-Based Unsupervised Training





⇒ benefit of confidence term is limited, margin again significant

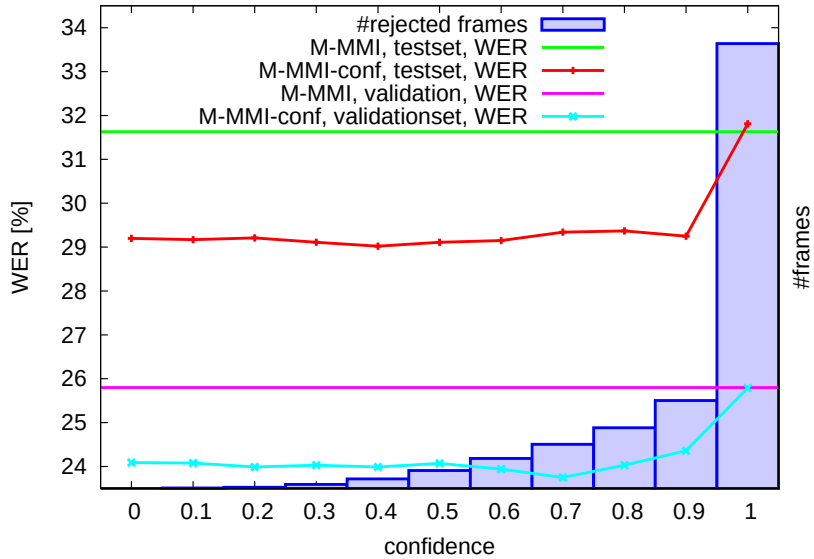
⇒ typically M-MMI/M-MMI-conf more robust

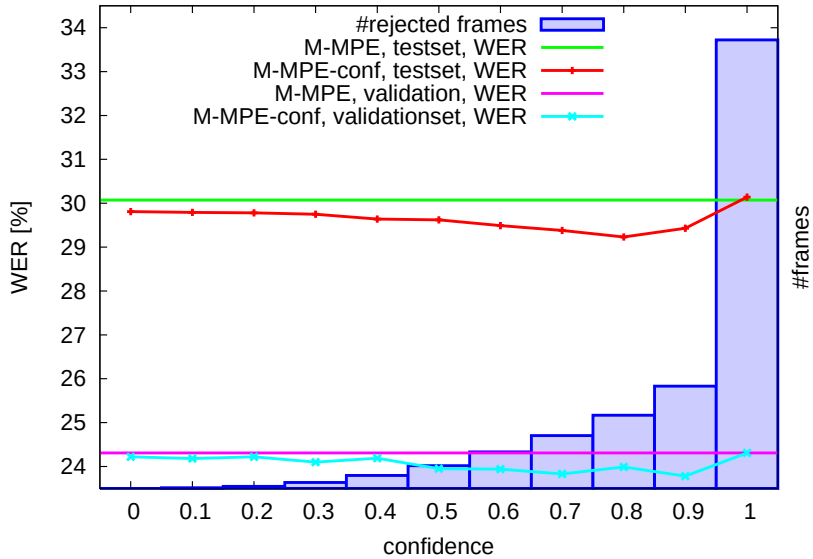
⇒ confidence term is important in M-MPE-conf

Continuous Latin Lines - IAM

Optical Character Recognition

M-MMI-conf vs. M-MPE-conf





⇒ word graph density is important for smooth convergence

⇒ typically M-MMI/M-MMI-conf more robust

⇒ M-MPE usually (slightly) outperforms M-MMI

⇒ benefit of margin term

Continuous Arabic Lines - RAMP-N

Optical Character Recognition

- Arabic machine-print
- open vocabulary
- 106k lexicon, 3-gram LM
- RWTH Arabic Machine-Print Newspaper (RAMP-N) corpus

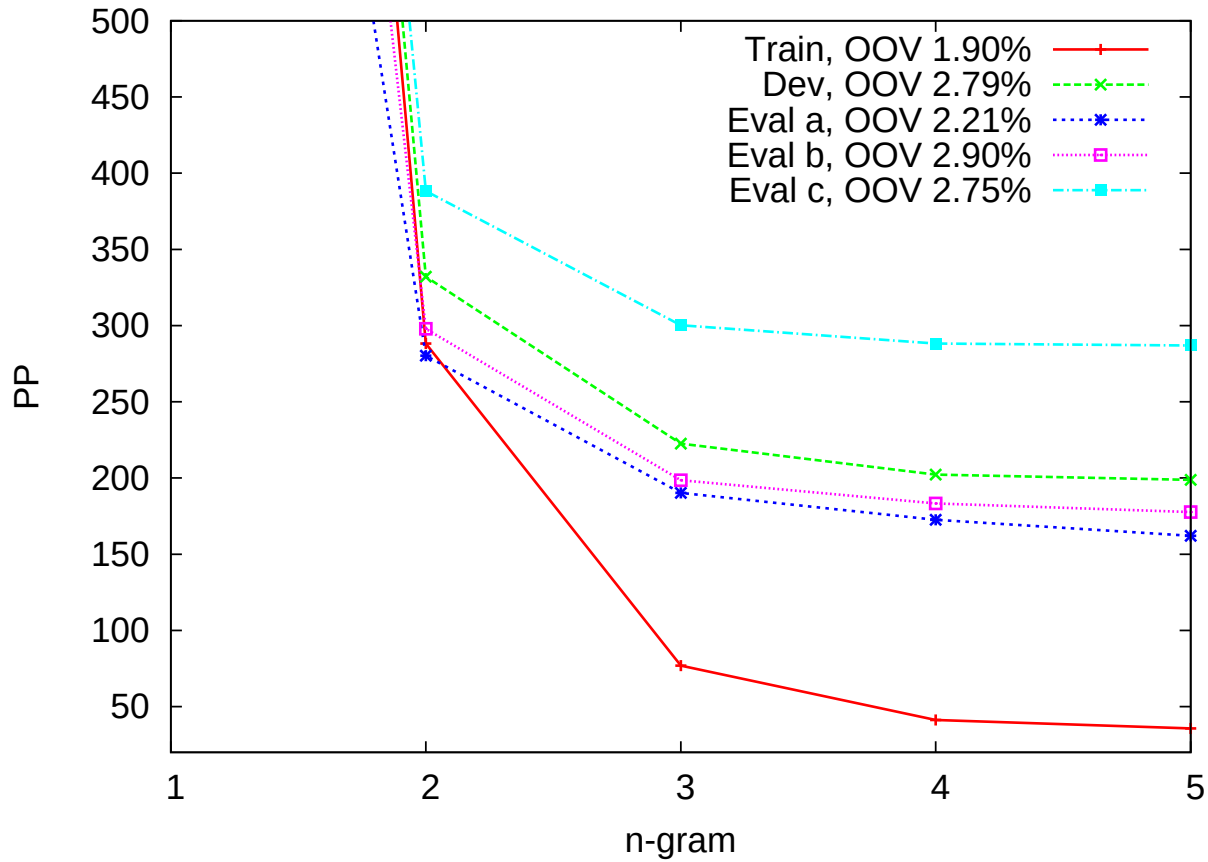
	Train	Dev	Eval a	Eval b	Eval c	LM Training
words	1.4M	7.7k	20.0k	17.2k	15.2k	228M
characters	5.9M	30.8k	72.3k	64.2k	62.0k	989M
lines	222.4k	1.1k	3.4k	2.4k	2.2k	22M
pages	409	2	5	4	4	85k
fonts	20	5	12	7	6	-
OOV rate	1.9%	2.8%	2.2%	2.9%	2.7%	5.5%

Return

Perplexities - RAMP-N

Optical Character Recognition

- LM using modified Kneser-Ney smoothing
- vocabulary size of 106k words



Return

Arabic Machine-Print - Ground-Truth

Optical Character Recognition

Why?

- there is no suitable OCR database

Goals?

- large-vocabulary OCR database
  $> \mathbf{1}$M training words
  $> \mathbf{200}$M language model ($\sim$ ASR)

How?

- PDF $\Rightarrow$ "Automatic" Ground-Truth
  - OCRopus and PDFlib TET

Arabic Machine-Print - Ground-Truth

Optical Character Recognition

Visual Model Data

- 450 PDFs, 1 Arabic Newspapers, May 2010 - Nov 2010
- total size: 247k lines, 1.6M words, 8.5M characters, 20 fonts

Language Model Data

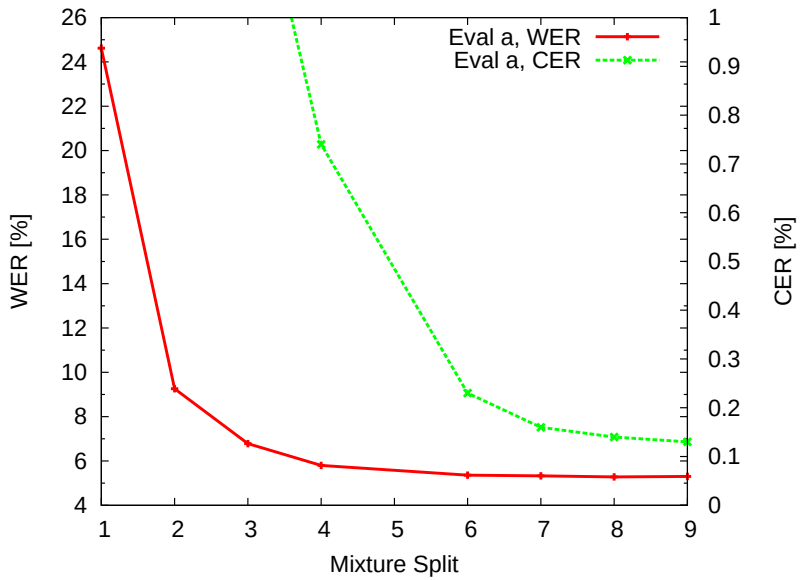
- 85k PDFs, 2 Arabic Newspapers, Jan 2003 - Aug 2010
- total size: 22M lines, 228M words, 989M characters

$\Rightarrow$ there is much more available ...

$\Rightarrow$ multi-font re-rendering possible ...

ML Trained GHMMs

- Eval a = 20k words, 72k chars, 3.4k lines, 12 fonts
- 106k lexicon ⇒ 2.2% OOVs, 3-gram word LM ⇒ 190 PPL

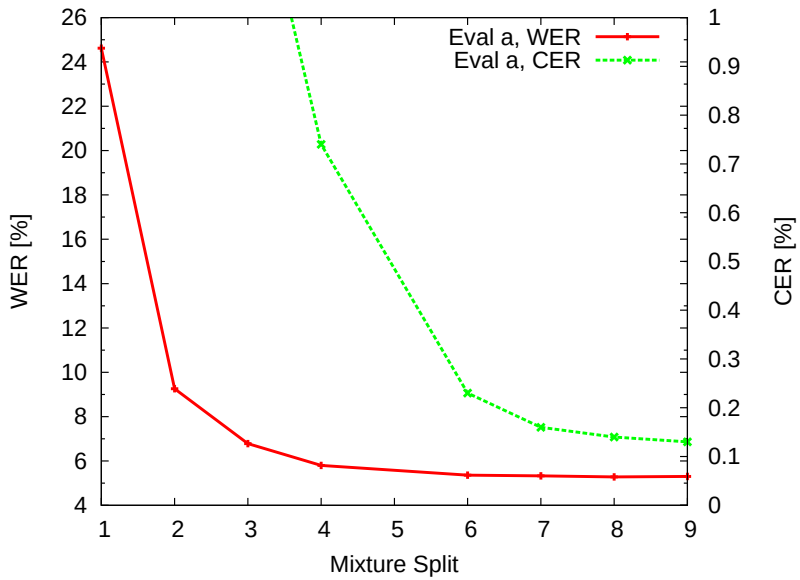


⇒ GMMs can cope with multiple-fonts

⇒ important with OOVs: focus on CER instead of WER

ML Trained GHMMs

- Eval a = 20k words, 72k chars, 3.4k lines, 12 fonts
- 106k lexicon ⇒ 2.2% OOVs, 3-gram word LM ⇒ 190 PPL



⇒ GMMs can cope with multiple-fonts

⇒ important with OOVs: focus on CER instead of WER

⇒ Character LMs (8-gram): 1.28 % CER

Continuous Arabic Lines - RAMP-N

Optical Character Recognition

ML trained GHMMs using Rendered and Scanned data

Layout Analysis	Rendered		Scanned	
	WER[%]	CER[%]	WER[%]	CER[%]
Supervised	4.76	0.15	5.79	0.64
Unsupervised	-	-	17.62	3.79

$\Rightarrow$ rendered data: remaining errors mainly due to OOVs

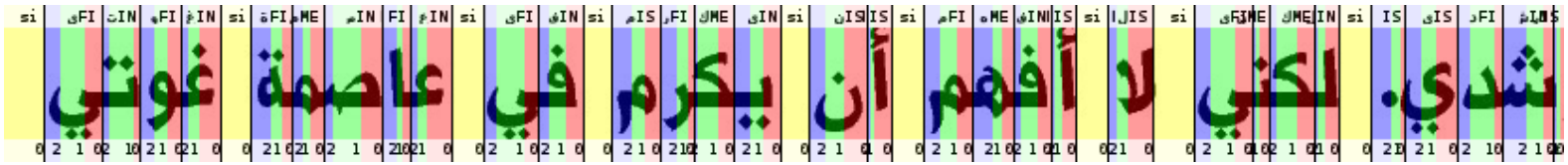
$\Rightarrow$ scanned data: problems with OCRopus and feature robustness

Continuous Arabic Lines - RAMP-N

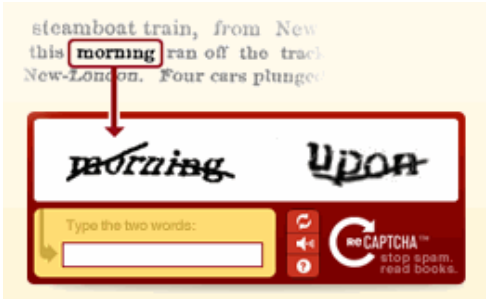
Optical Character Recognition

Visual Inspection

- ML



⇒ implicit HMM segmentation adequate for post-processing



ML trained GHMMs

- Font-dependent results on the RAMP-N subset Eval a

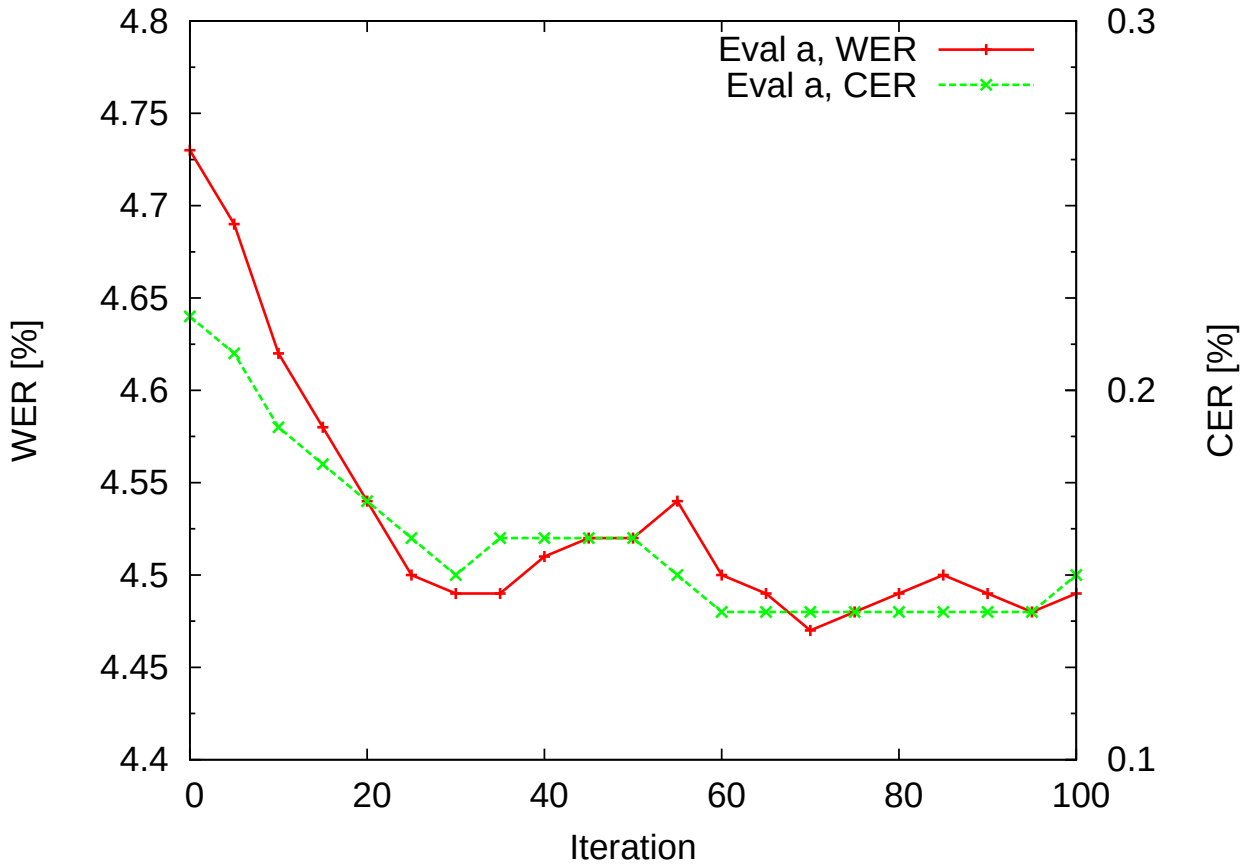
Font	Lines	Errors	Words	OOV	WER[%]	Errors	Glyphs	CER[%]
AXtAlFares	2	10	2	2	**500.0**	0	19	**0.00**
AXtCalligraph	1	0	8	0	0.0	0	21	0.00
AXtGIHaneBoldItalic	15	19	129	4	14.7	12	591	2.03
AXtHammed	3	0	5	0	0.0	0	31	0.00
AXtKaram	9	2	83	0	2.4	4	300	1.33
AXtManal	1	0	2	0	0.0	0	4	0.00
AXtManalBlack	5	5	27	1	18.5	11	112	9.82
AXtMarwanBold	109	46	385	18	**12.0**	13	2002	**0.65**
AXtMarwanLight	3261	828	18963	405	**4.4**	79	83091	**0.10**
AXtShareQ	5	10	64	0	15.6	7	299	2.34
AXtShareQXL	68	35	371	13	**9.4**	10	1973	**0.51**
AXtThuluthMubassat	1	0	3	0	0.0	0	13	0.00
Total (Eval a)	3480	955	20042	443	4.8	136	88456	0.15

◂ Return

Continuous Arabic Lines - RAMP-N

Optical Character Recognition

M-MPE Trained GHMMs



$\Rightarrow$ remaining errors mainly due to OOVs

Appendix: Font Examples - RAMP-N

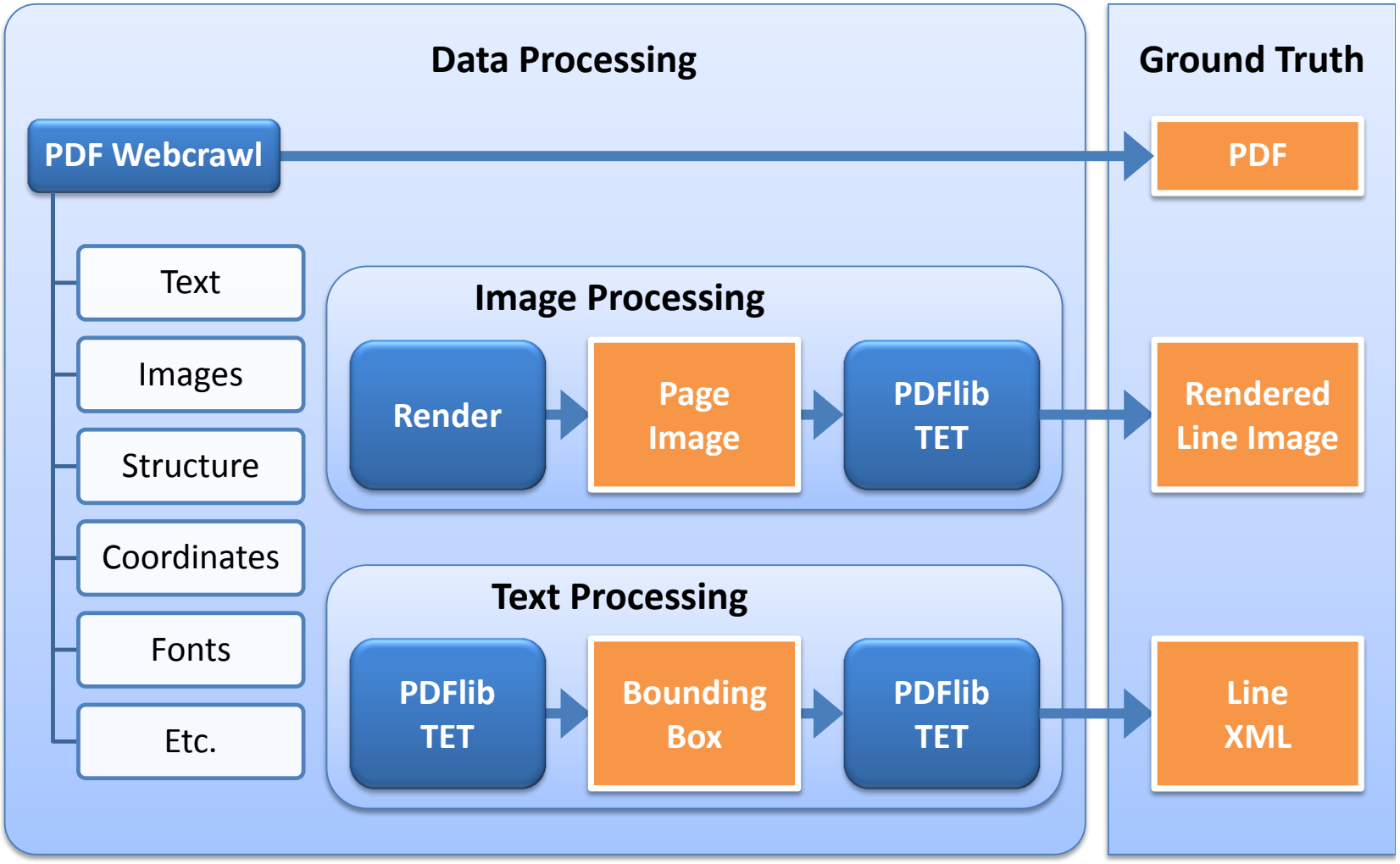
Experimental Results

AXtAlFAres	الخط الحسن يزيد الحق وضوحا
AXtCalligraph	الخط الحسن يزيد الحق وضوحا
AXtGlHaneBoldItalic	الخط الحسن يزيد الحق وضوحا
AXtHammed	الخط الحسن يزيد الحق وضوحا
AXtKarim	الخط الحسن يزيد الحق وضوحا
AXtManalFont	الخط الحسن يزيد الحق وضوحا
AXtMarwanBold	الخط الحسن يزيد الحق وضوحا
AXtMarwanLight	الخط الحسن يزيد الحق وضوحا
AXtSHAReQ	الخط الحسن يزيد الحق وضوحا
AXtSHAReQXL	الخط الحسن يزيد الحق وضوحا
AXtThuluthMubassat	الخط الحسن يزيد الحق وضوحا

Ground-Truth - RAMP-N

Optical Character Recognition

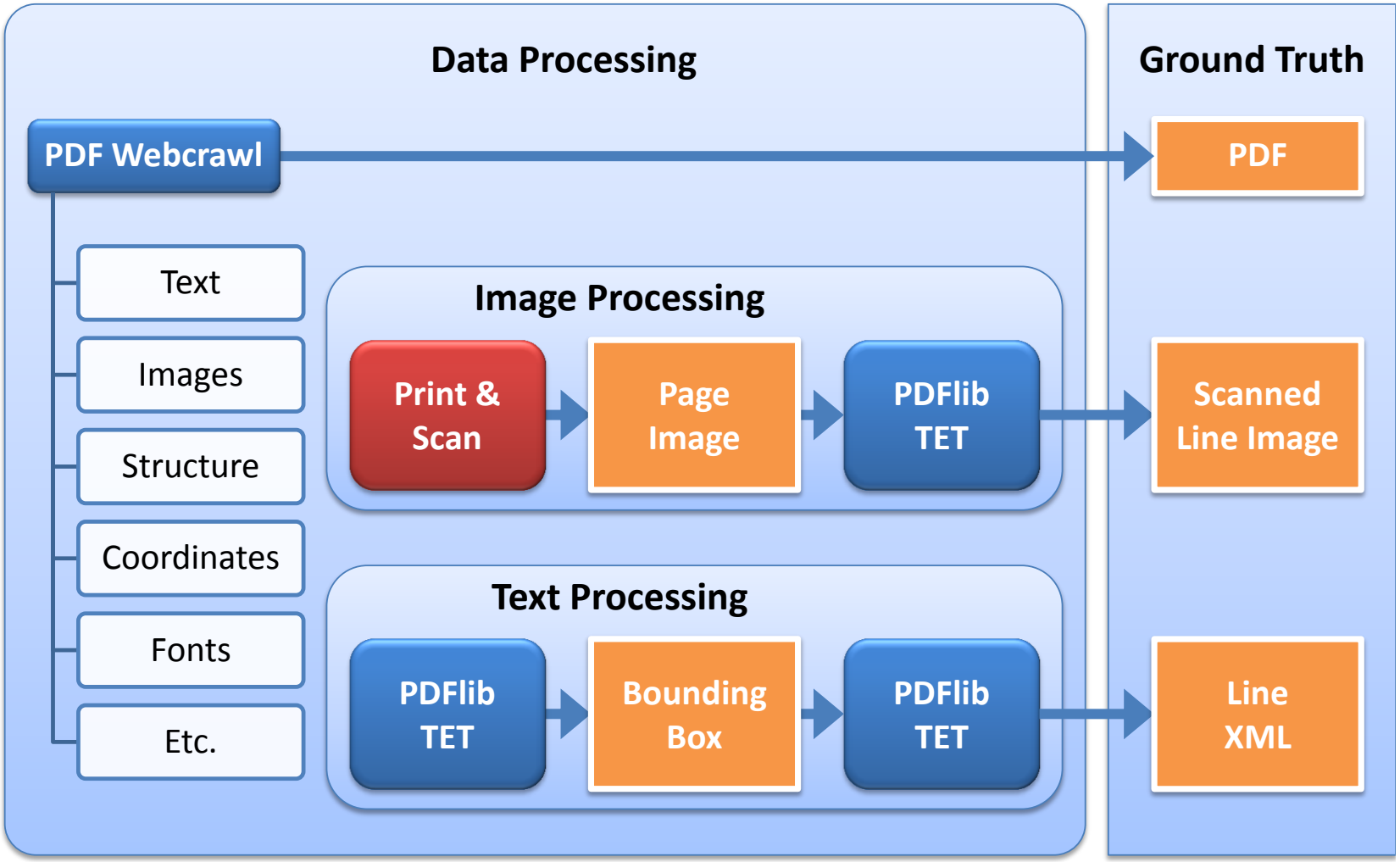
Supervised



Ground-Truth - RAMP-N

Optical Character Recognition

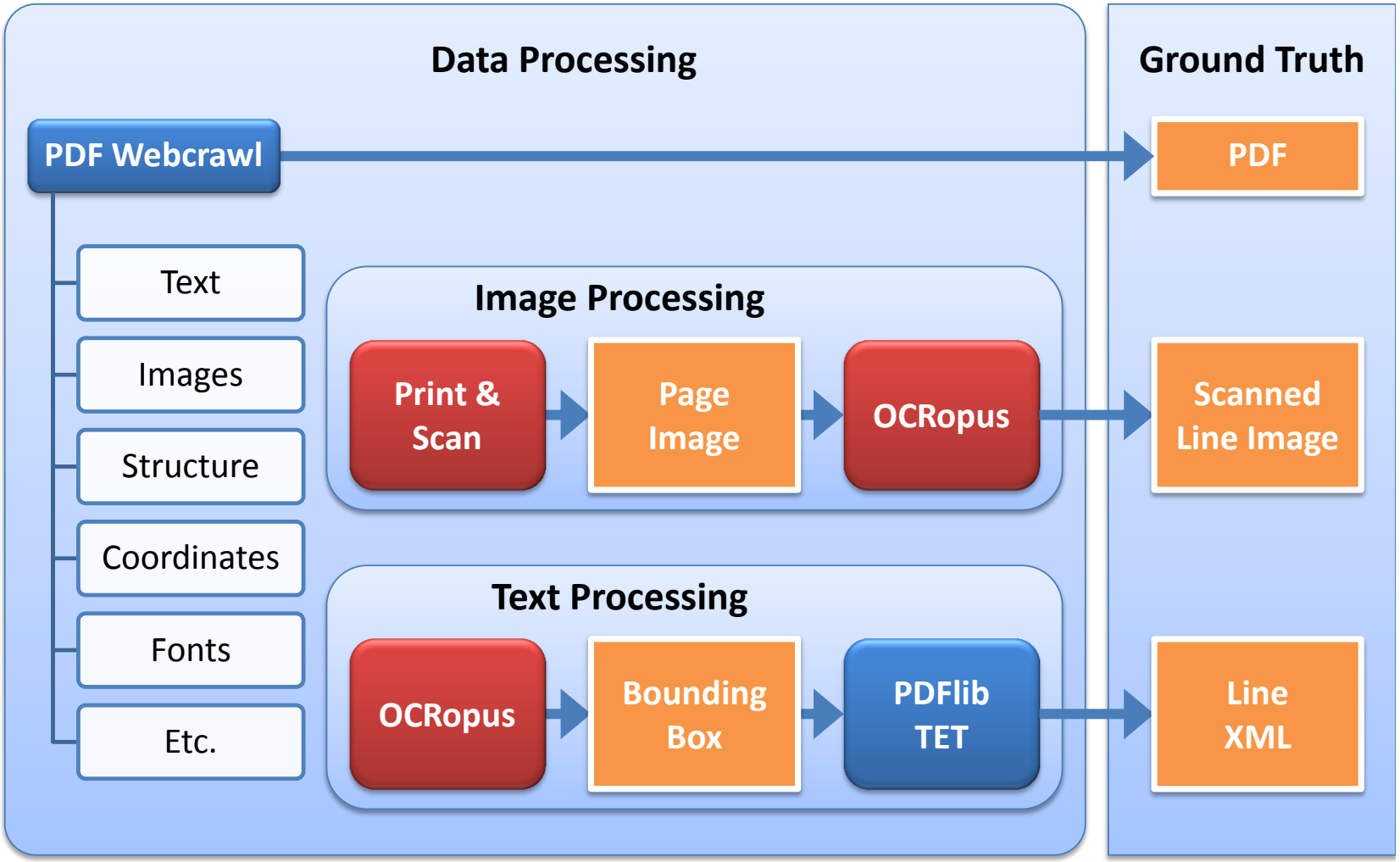
Semi-Supervised



Ground-Truth - RAMP-N

Optical Character Recognition

Unsupervised



Appendix: Automatic Sign Language Recognition

RWTH AACHEN UNIVERSITY

Introduction

Automatic Sign Language Recognition

Problems to be Solved in ASR/ASLR

1. preprocessing and feature extraction of the input signal
2. specification of models for the words to be recognized
3. learning of the free model parameters from the training data
4. maximum probability search over all models during recognition

Similarities

- temporal sequence of sounds or gestures
- languages and dialects

Main Differences Between Signed and Spoken Languages

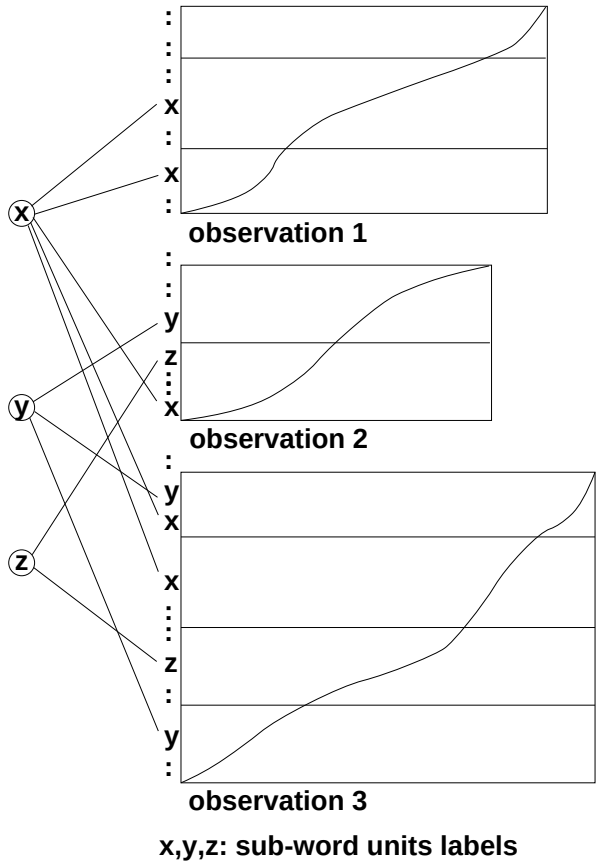
- simultaneousness
- signing space
- 3D coarticulation and movement epenthesis
- silence

Feature Extraction and Modeling

Automatic Sign Language Recognition

Sub-Word Units

- possible to recognize unseen words using a pronunciation lexicon
- problems in sign language recognition:
  - phoneme still not well-defined
  - phonemes occur simultaneously
  - no unique pronunciation lexicon
  - more phonemes in sign language



$\Rightarrow$ approach not directly transferable to sign language recognition

$\Rightarrow$ usually whole-word models are used

$\Rightarrow$ 3-state HMM, GMMs