Measuring HMM similarity with the Bayes probability of error

An application to on-line handwritten character recognition

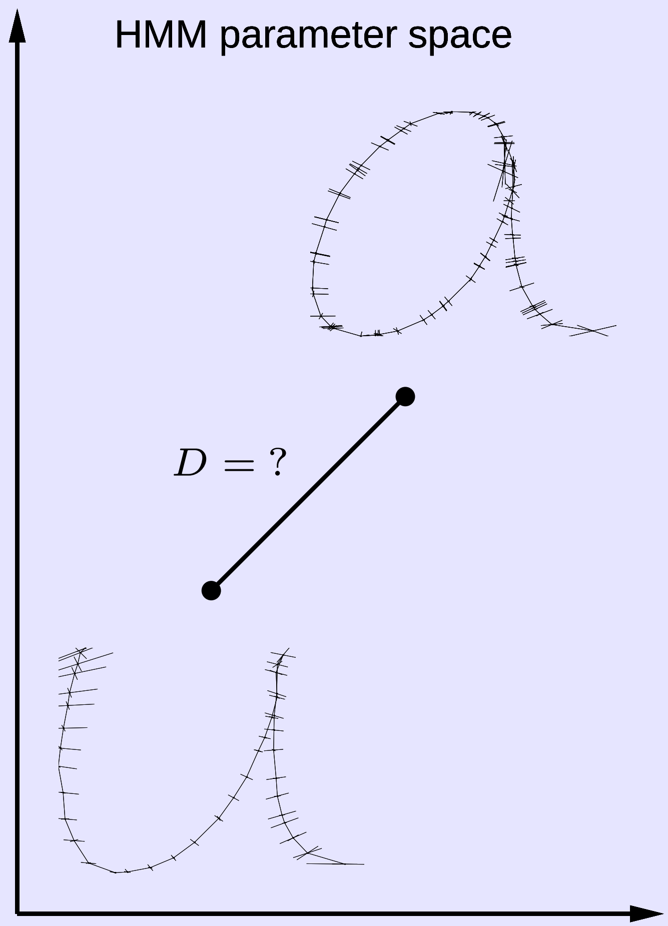
Dipl.-Inf. Claus Bahlmann

Abstract

- Introduction
- Sequence Classifiers
  - Dynamic Time Warping (DTW)
  - Statistical DTW/HMM
- Proposed HMM Similarity Measure
- Experiments with on-line handwritten characters

Introduction

Problem



Context

On-line handwritten character recognition

Why do we need HMM distance?

- detection of “close” competing HMMs
- interpretation of misclassifications
- monitoring iterative training process
- HMM clustering

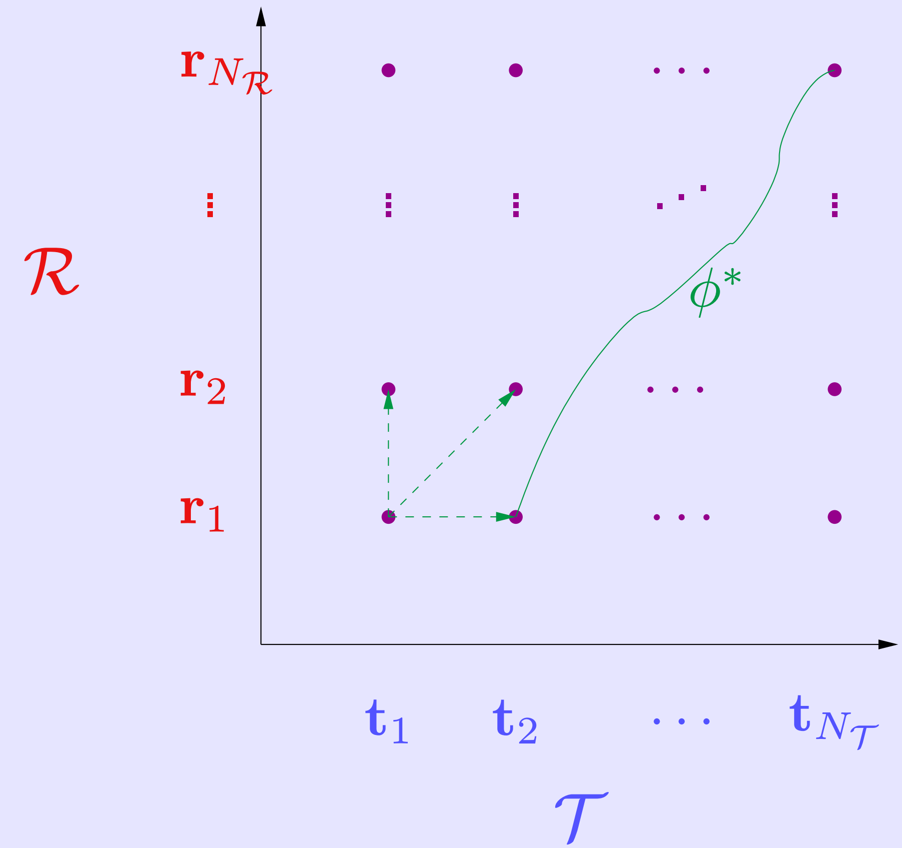
Approach

classification-oriented

Dynamic Time Warping (DTW)

Alignment distance:

$$D_{\phi}(\mathcal{T},\mathcal{R}) = \frac{1}{N}\sum_{i=1}^{N} d\left(\mathbf{t}_{\phi_{\mathcal{T}(i)}}, \mathbf{r}_{\phi_{\mathcal{R}(i)}}\right)$$



Viterbi distance:

$$D(\mathcal{T},\mathcal{R}) = D_{\phi^*}(\mathcal{T},\mathcal{R}) = \min_{\phi}\{D_{\phi}(\mathcal{T},\mathcal{R})\}$$

Local distance: Euclidean distance

$$d(\mathbf{t}_i,\mathbf{r}_j) = \|\mathbf{t}_i - \mathbf{r}_j\|$$

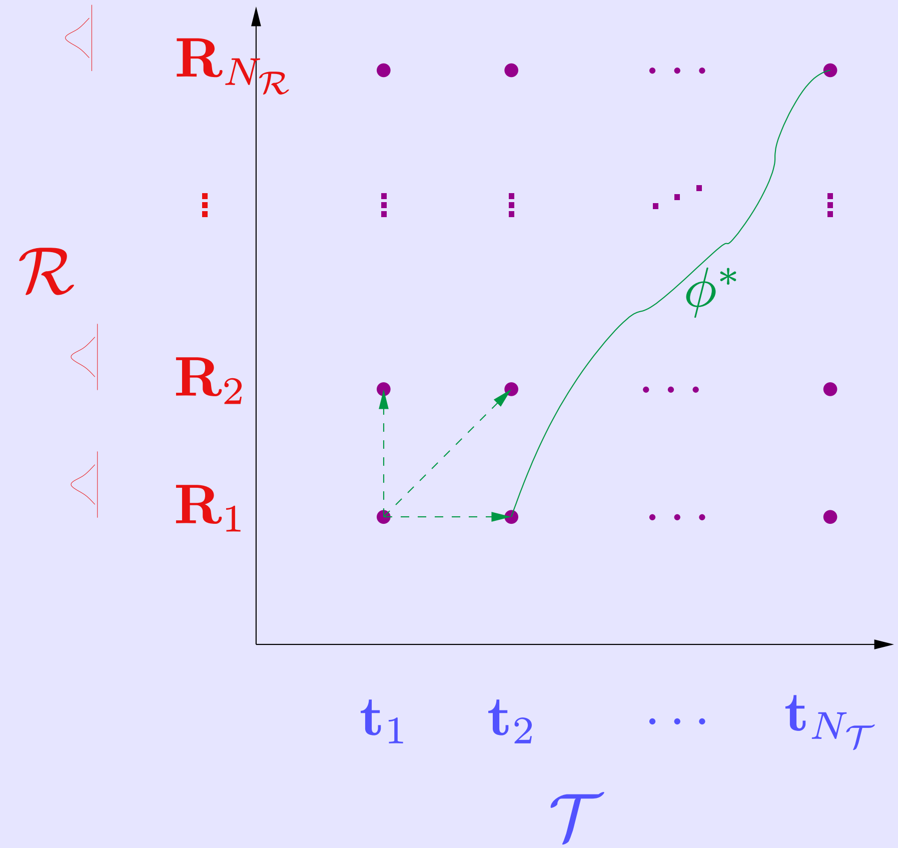
Statistical DTW (SDTW), HMM

Alignment distance:

$$D_{\phi}\left(\mathcal{T},\mathcal{R}\right)=\frac{1}{N}\sum_{i=1}^{N}d\left(\mathbf{t}_{\phi_{\mathcal{T}(i)}},\mathbf{R}_{\phi_{\mathcal{R}(i)}}\right)$$

Viterbi distance:

$$D\left(\mathcal{T},\mathcal{R}\right)=D_{\phi^*}\left(\mathcal{T},\mathcal{R}\right)=\min_{\phi}\left\{D_{\phi}\left(\mathcal{T},\mathcal{R}\right)\right\}$$



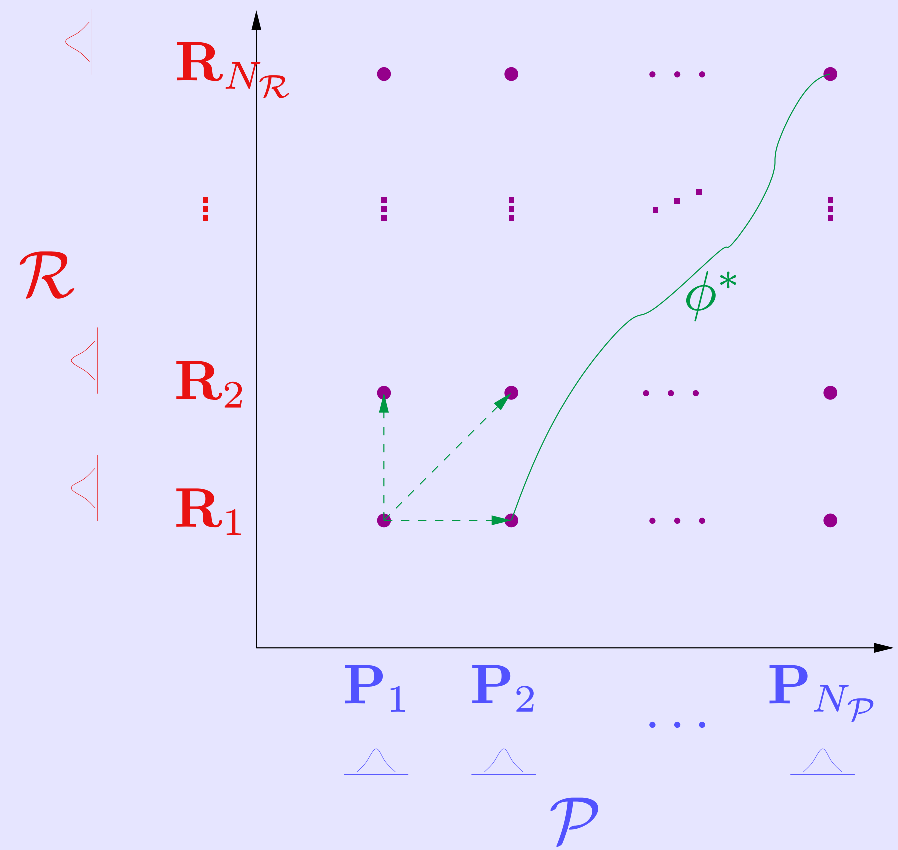
Local distance: $-\log$ a-posteriori probability

$$d\left(\mathbf{t}_i,\mathbf{R}_j\right)=-\log P\left(\mathbf{t}_i|\mathbf{R}_j\right)$$

SDTW/HMM Similarities

Alignment distance:

$$D_{\phi}\left(\mathcal{P},\mathcal{R}\right) = \frac{1}{N}\sum_{i=1}^{N} d\left(\mathbf{P}_{\phi_{\mathcal{T}(i)}},\mathbf{R}_{\phi_{\mathcal{R}(i)}}\right)$$



Viterbi distance:

$$D\left(\mathcal{P},\mathcal{R}\right) = D_{\phi^*}\left(\mathcal{P},\mathcal{R}\right) = \min_{\phi}\left\{D_{\phi}\left(\mathcal{P},\mathcal{R}\right)\right\}$$

Local distance:

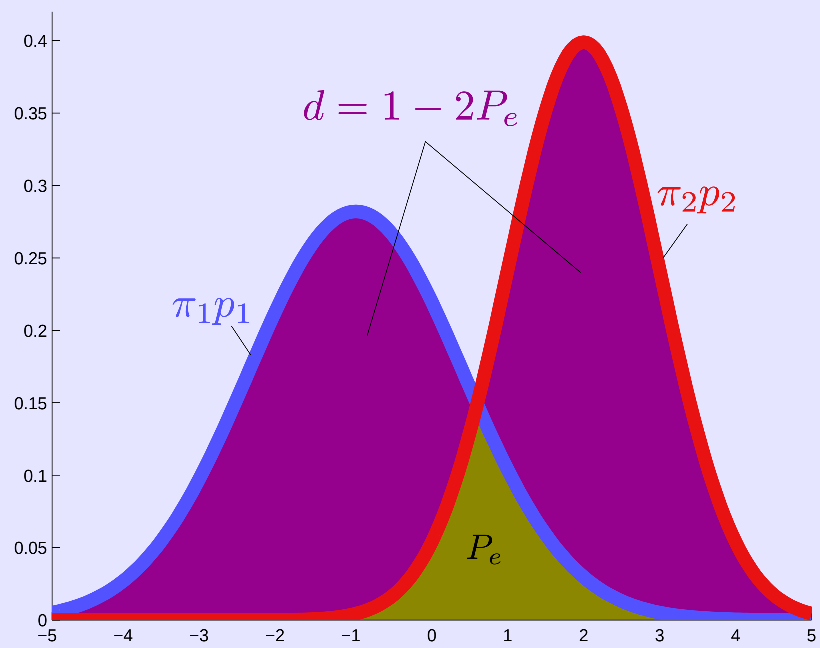
$$d\left(\mathbf{P}_i,\mathbf{R}_j\right) \quad = \quad ?$$

Bayes Error

Overlap of two pdfs p_1 and p_2 with prior probabilities π_1 and π_1.

$$P_e\left(p_1\left(\mathbf{x}\right), p_2\left(\mathbf{x}\right)\right) = \int_{\mathbf{x}} \min\left\{\pi_1 p_1\left(\mathbf{x}\right), \pi_2 p_2\left(\mathbf{x}\right)\right\} d\mathbf{x}$$

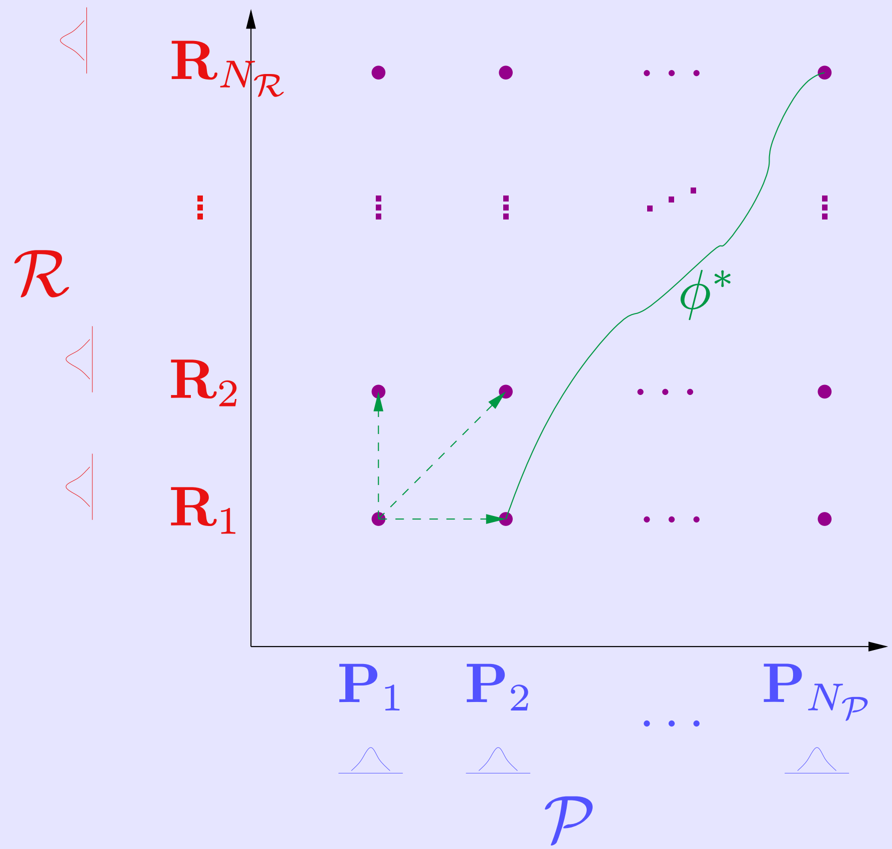
$$d\left(p_1\left(\mathbf{x}\right), p_2\left(\mathbf{x}\right)\right) = 1 - 2P_e\left(p_1\left(\mathbf{x}\right), p_2\left(\mathbf{x}\right)\right)$$



SDTW/HMM Similarities

Alignment distance:

$$D_{\phi}\left(\mathcal{P},\mathcal{R}\right)=\frac{1}{N}\sum_{i=1}^{N} d\left(\mathbf{P}_{\phi_{\mathcal{T}(i)}},\mathbf{R}_{\phi_{\mathcal{R}(i)}}\right)$$



Viterbi distance:

$$D\left(\mathcal{P},\mathcal{R}\right)=D_{\phi^*}\left(\mathcal{P},\mathcal{R}\right)=\min_{\phi}\left\{D_{\phi}\left(\mathcal{P},\mathcal{R}\right)\right\}$$

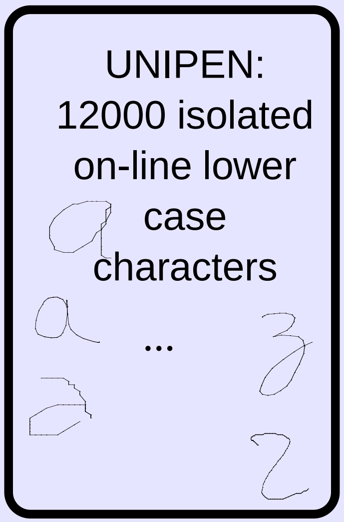
Local distance:

$$d\left(\mathbf{P}_i,\mathbf{R}_j\right) \quad = \quad 1-2P_e\left(\mathcal{N}\left(\mathbf{P}_i,\mathbf{x}\right),\mathcal{N}\left(\mathbf{R}_j,\mathbf{x}\right)\right)$$

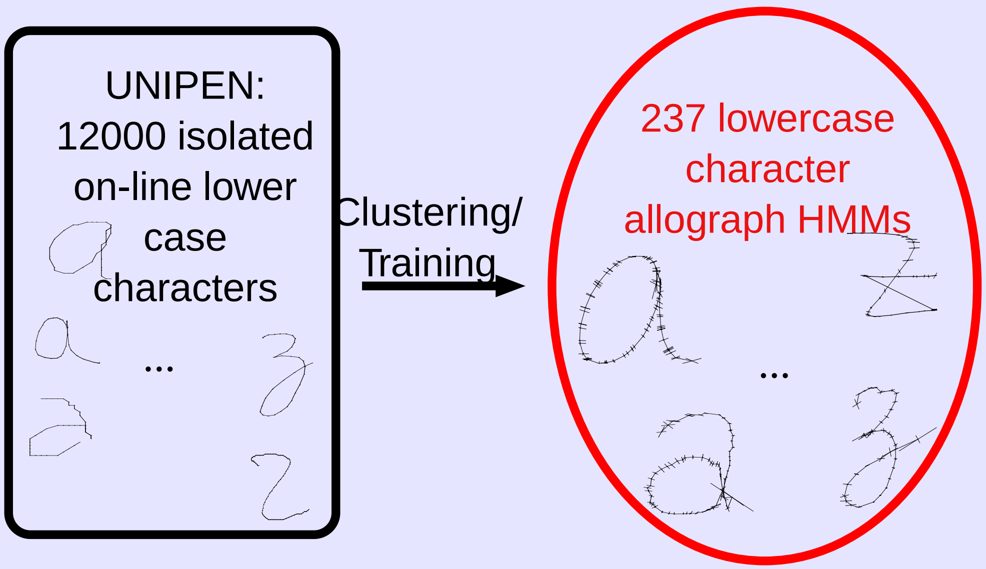
Backtransformation:

$$P_e^*\left(\mathcal{P},\mathcal{R}\right)=\frac{1}{2}\left(1-D\left(\mathcal{P},\mathcal{R}\right)\right)$$

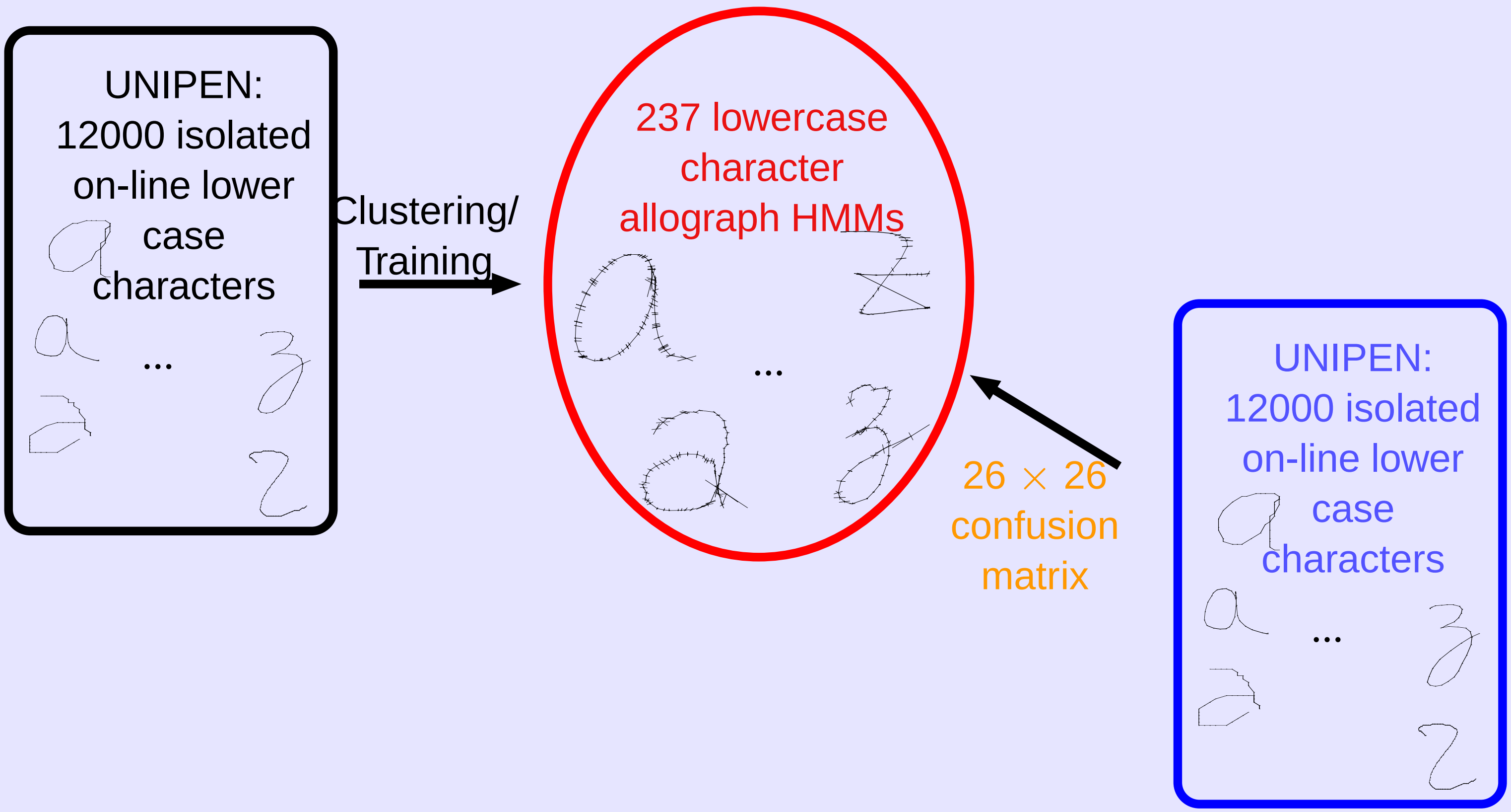
Description of experiments



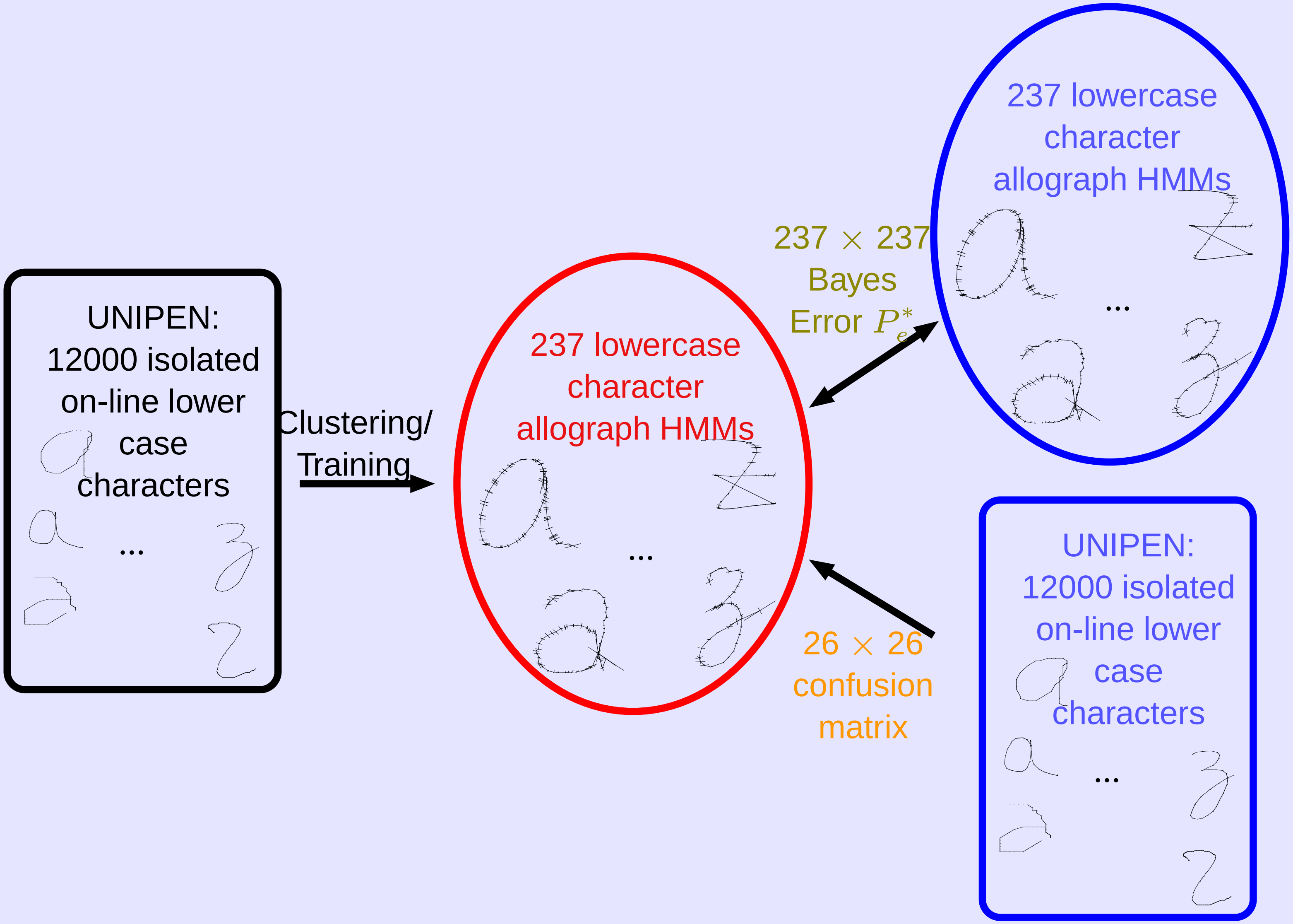
Description of experiments



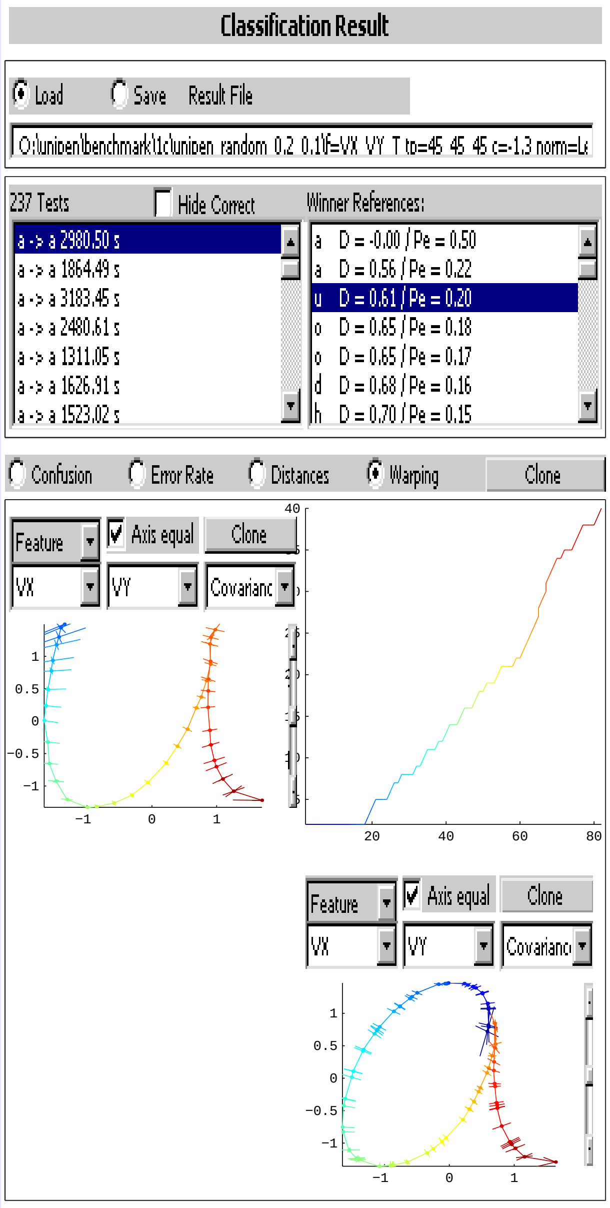
Description of experiments



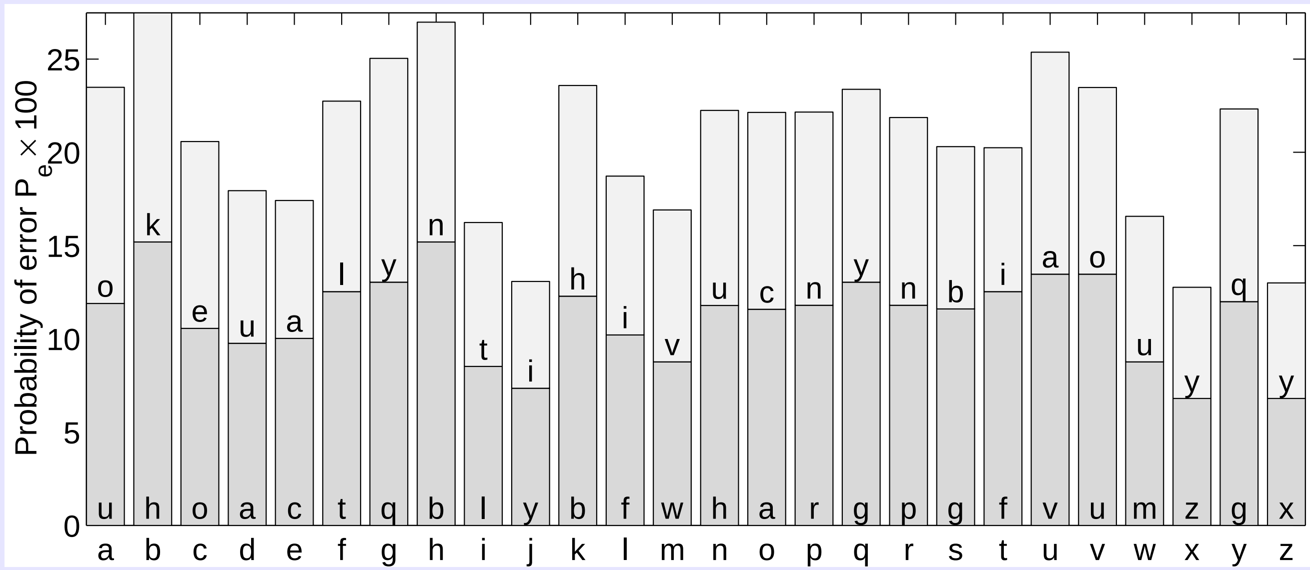
Description of experiments

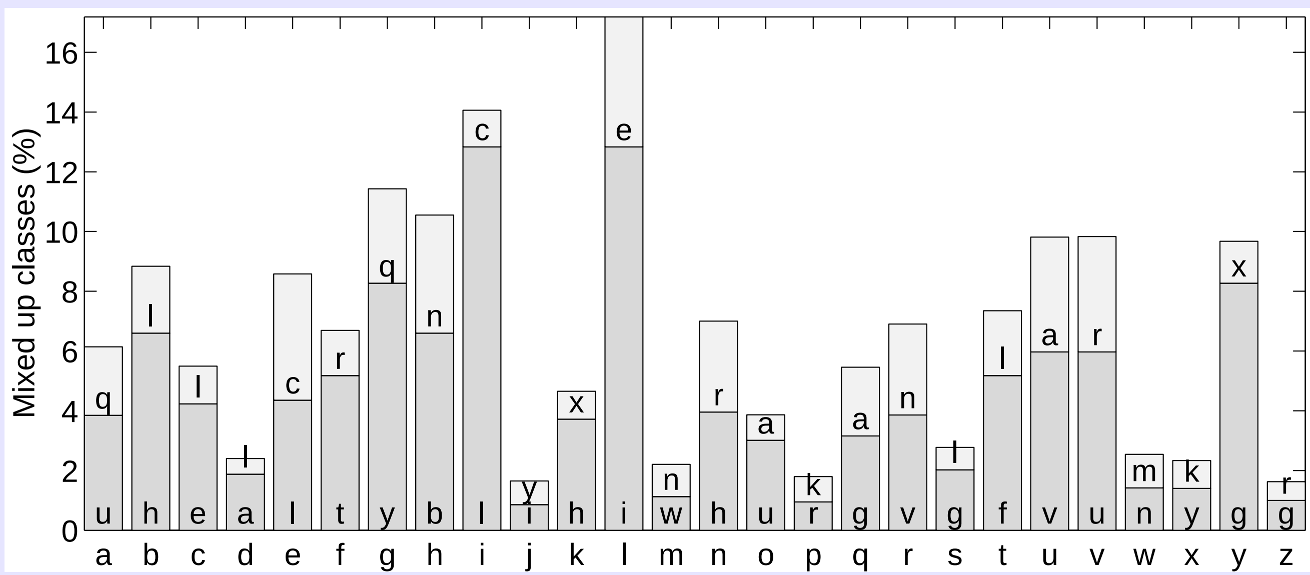


Examples of Bayes error



Comparing Bayes Error and Classification Confusions





Conclusion

- Several Applications for distance measure
- Introduced HMM Similarity Measure
  - DTW / HMM classification as starting point
  - flexible in definition of local distance (Bayes error, χ^2, Kullback-Leibler or Jensen-Shannon.)
  - not limited to handwriting recognition
- Experiments show
  - correspondence of similarity measure with visually assessed similarity
  - qualitative correlation of most similar and mostly confused classes

Future Work

- Discriminative training / hybrid classifiers for *similar* HMMs
- distance measure as stop criterion for iterative HMM training
- modeling transition probabilities
- Gaussian mixture models

Comparing Error Probability and Misclassifications

$P_e^*\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)$: Error probability of prototype $l'k'$ and lk

$C_{l',l}$: Number of classifications from class l into l'

Prob. of error	Misclassification
$237^2 \times P_e^*\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)$	$26^2 \times C_{l'l}$
$\downarrow$	$\downarrow$
$\tilde{P}_e^*(l', l) = \mathcal{E}\left[P_e^*\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)\right]_{k',k}$	$C'_{l'l} = C_{l'l} / \left(C_{l'l} + C_{ll}\right)$
	$\downarrow$
	$\tilde{C}_{l'l} = \tilde{\pi}_{l'l} C'_{l'l} + \tilde{\pi}_{ll'} C'_{ll'}$

Feature Extraction

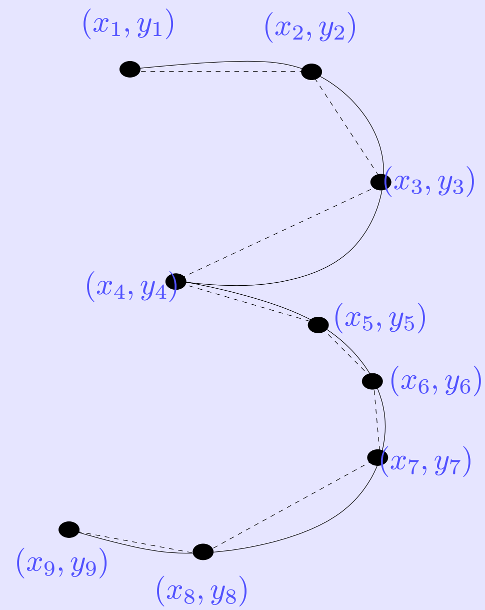
- Data:

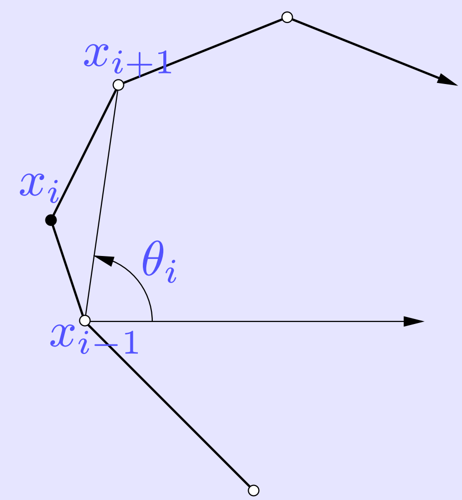
  polygon $[(x_i, y_i)]_{i=1,\ldots,N_{\mathcal{T}}}$

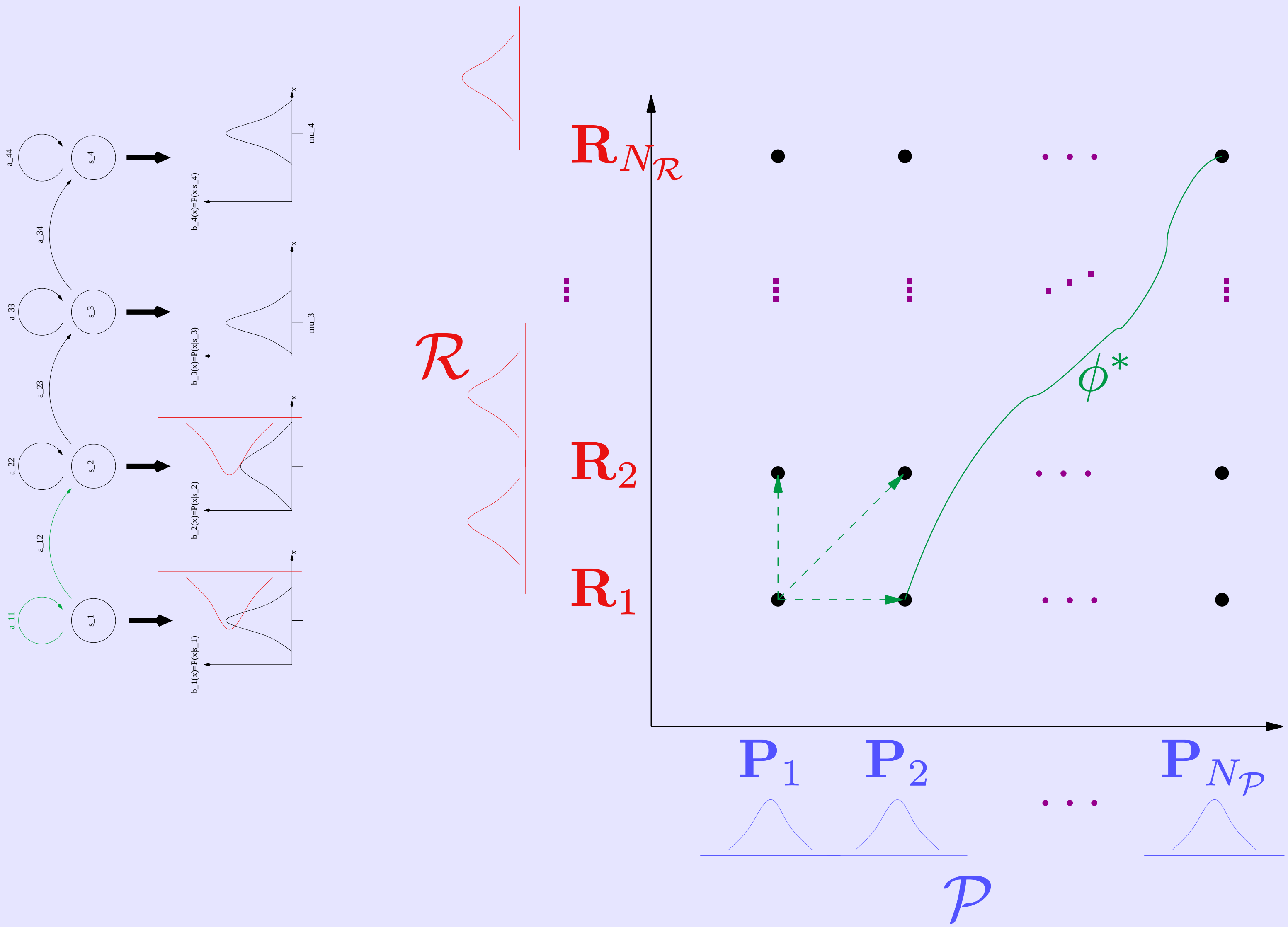
- Features:
  - normalized x-coordinate $\tilde{x}_i = \frac{x_i - \mu_x}{\sigma_y}$
  - normalized y-coordinate $\tilde{y}_i = \frac{y_i - \mu_y}{\sigma_y}$
  - tangent angle $\theta_i = \text{ang}\left((x_{i+1} - x_{i-1}) + \jmath \cdot (y_{i+1} - y_{i-1})\right)$

$$\text{feature vector } \mathbf{t}_i = (\tilde{x}_i, \tilde{y}_i, \theta_i)^T$$

$$\text{writing } \mathcal{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_{N_{\mathcal{T}}})$$







$\mathbf{R}_{N_{\mathcal{R}}}$
$\mathcal{R}$
$\mathbf{R}_2$
$\mathbf{R}_1$
ϕ^*
$\mathbf{P}_1$
$\mathbf{P}_2$
$\mathbf{P}_{N_{\mathcal{P}}}$
$\mathcal{P}$