publ. in Proc. of the 6th Int. Conf. on Document Analysis and Recognition (ICDAR) 2001, pp. 406–411, 2001.
ftp://ftp.informatik.uni-freiburg.de/papers/lmb/ba_bu_icdar01.pdf

Measuring HMM Similarity with the Bayes Probability of Error and its Application to Online Handwriting Recognition

Claus Bahlmann and Hans Burkhardt
Computer Science Department
Albert Ludwigs University Freiburg
79110 Freiburg, Germany
{bahlmann,burkhardt}@informatik.uni-freiburg.de

Abstract

We propose a novel similarity measure for Hidden Markov Models (HMMs). This measure calculates the Bayes probability of error for HMM state correspondences and propagates it along the Viterbi path in a similar way to the HMM Viterbi scoring. It can be applied as a tool to interpret misclassifications, as a stop criterion in iterative HMM training or as a distance measure for HMM clustering. The similarity measure is evaluated in the context of online handwriting recognition on lower case character models which have been trained from the UNIPEN database. We compare the similarities with experimental classifications. The results show that similar and misclassified class pairs are highly correlated. The measure is not limited to handwriting recognition, but can be used in other applications that use HMM based methods.

1 Introduction

Hidden Markov Model (HMM) based techniques are frequently used for modeling handwritten symbols like characters or sub-characters in handwriting recognition. Their success is based mainly on powerful methods to adapt and score them. The complex structure of HMMs makes it difficult however to give a unique measure of similarity between two (or more) HMMs. In this contribution we propose a method for comparing HMMs based on the Bayes probability of error. We also show experimentally gained measures for lower case character models which are trained on the UNIPEN online handwriting database [3] and compare them to classification results.

Applications of HMM similarity measures are manifold. They can be used for monitoring and controlling the parameter re-estimation during the training process or as a measure for discriminative training methods [6] as well as for HMM clustering. Furthermore, and this is the objective in this contribution, a similarity measure can help get a thorough insight into misclassifications.

HMM similarity or dissimilarity measures have been proposed by a few authors. Early approaches were based on the Euclidean distance of the discrete observation probabilities [7], others on entropy [4, 2] or on co-emission probabilities [9] of two models.

The proposed measure is defined on the basis of the Bayes probability of error and thus is well suited for analyzing misclassification behavior. It integrates very elegantly into common HMM based classification methods, as will be demonstrated. We shall therefore start with a short review of the underlying handwriting recognition system in the following section. Section 3 then describes the similarity measure and experimental results with this measure are presented in section 4. Section 5 provides a summary of this contribution.

2 Recognition

2.1 Feature selection

Here we are dealing with a system for online data. A character is represented as a polygon $\mathcal{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_{N_T})$. Each element $\mathbf{t}_i$ describes a feature vector at sample point i. In our case, the feature vector is $\mathbf{t}_i = (\tilde{x}_i, \tilde{y}_i, \theta_i)^T$: $\tilde{x}_i = \frac{x_i - \mu_x}{\sigma_y}$ and $\tilde{y}_i = \frac{y_i - \mu_y}{\sigma_y}$ are the horizontal x- and vertical y-coordinates of the writing shifted by the character mean (μ_x, μ_y) and normalized by the y-deviation σ_y. The feature θ_i is the tangent slope angle at point i, approximated by $\theta_i = \text{ang}\left((x_{i+1} - x_{i-1}) + \jmath \cdot (y_{i+1} - y_{i-1})\right)$ with $\jmath^2 = -1$ and "ang" the complex angle function.

The data is sampled at regular interval in time. No preprocessing, such as re-sampling of the writing or reference

line detection, is applied in this case. Each pattern is typically represented by about 30–80 samples.

2.2 Classification

Classification is accomplished by dynamic time warping (DTW) [11] with an extension to incorporate a second order statistic, a technique which we call "statistical DTW" (SDTW).

To be more specific: the classifier is defined by the minimum distance $D\left(\mathcal{T},\mathcal{R}^{lk}\right)$ of the test pattern $\mathcal{T}$ to a set of reference models $\mathcal{R}^{lk}$ and the prior probabilities π_{lk} (l corresponds to a character class and k to an allograph prototype, i.e. a characteristic shape of class l)

$$\widehat{l} = \arg \min_{l\in\{1,\ldots,L\},k\in\{1,\ldots,K_l\}} \left\{D\left(\mathcal{T},\mathcal{R}^{lk}\right) - \log \pi_{lk}\right\}. \tag{1}$$

Given a warping (or alignment) path $\boldsymbol{\phi} = (\phi(1),\ldots,\phi(N))$ with $\phi(i) = (\phi_{\mathcal{T}}(i),\phi_{\mathcal{R}}(i))$ (see [11, Chapter 4.7] for further details), $D_{\boldsymbol{\phi}}$ is defined as the normalized, accumulated distance along the warping path

$$D_{\boldsymbol{\phi}}\left(\mathcal{T},\mathcal{R}^{lk}\right) = \frac{1}{N}\sum_{i=1}^{N} d\left(\mathbf{t}_{\phi_{\mathcal{T}}(i)},\mathbf{R}^{lk}_{\phi_{\mathcal{R}}(i)}\right) \tag{2}$$

and the classifier distance $D\left(\mathcal{T},\mathcal{R}^{lk}\right)$ in (1) is the distance (2) along the Viterbi path $\boldsymbol{\phi}^*$

$$D\left(\mathcal{T},\mathcal{R}^{lk}\right) = D_{\boldsymbol{\phi}^*}\left(\mathcal{T},\mathcal{R}^{lk}\right) = \min_{\boldsymbol{\phi}}\left\{D_{\boldsymbol{\phi}}\left(\mathcal{T},\mathcal{R}^{lk}\right)\right\}. \tag{3}$$

Several choices of local continuity constraints which define the allowable warping paths are conceivable. We use the ones illustrated by the dashed lines in figure 1 (a)–(c).

In contrast to the simple DTW classification, our local distance is not simply the Euclidean distance $d\left(\mathbf{t}_i,\mathbf{r}^{lk}_j\right) = \left\|\mathbf{t}_i - \mathbf{r}^{lk}_j\right\|$ of two feature vectors $\mathbf{t}_i$ and $\mathbf{r}^{lk}_j$, but is computed from $\mathbf{t}_i$ and a second order statistic of estimated mean and covariance $\mathbf{R}^{lk}_j = \left(\boldsymbol{\mu}^{lk}_j,\boldsymbol{\Sigma}^{lk}_j\right)$ of samples from a reference model (see figure 1 (b)). Assuming that the probability density function (pdf) $p^{lk}_j(\mathbf{x})$ at sample point j can be modeled by a unimodal multivariate Gaussian $\mathcal{N}\left(\mathbf{R}^{lk}_j,\mathbf{x}\right)$, a formula for the local distance can be derived to

$$\begin{aligned} d\left(\mathbf{t}_i,\mathbf{R}^{lk}_j\right) &= \frac{n}{2}\ln(2\pi) + \frac{1}{2}\ln\left(\left|\boldsymbol{\Sigma}^{lk}_j\right|\right) \\ &\quad + \frac{1}{2}\left(\mathbf{t}_i - \boldsymbol{\mu}^{lk}_j\right)^T\left(\boldsymbol{\Sigma}^{lk}_j\right)^{-1}\left(\mathbf{t}_i - \boldsymbol{\mu}^{lk}_j\right) \\ &\quad - \ln\left(P\left(\phi(j)\,|\phi(j-1),l,k\right)\right) \end{aligned} \tag{4}$$

The variable n is the dimension of the feature space, i.e. $n = 3$ in our case. The framework of equations (1)–(4) can be derived from a *maximum-a-posteriori* (MAP) classifier approach assuming sequential independency of the pdfs and some other commonly made prerequisites [1].

Usual dynamic programming and beam search strategies [11] are applied to reduce the computational complexity when minimizing equation (3). We would like to state that the supplementary normalization with respect to the length of the path (N) in equation (2) is not considered in the optimization criterion, and thus the Viterbi path is a suboptimal solution in general. However, the classification based on this suboptimal Viterbi detection performs better in practice.

It is helpful to see the SDTW in relation to the standard HMM context [11]: the reference sample points j correspond to the HMM states, the statistic $\mathbf{R}^{lk}_j = \left(\boldsymbol{\mu}^{lk}_j,\boldsymbol{\Sigma}^{lk}_j\right)$ corresponds to the observation pdf which is associated with the HMM state j and $P\left(\phi(j)\,|\phi(j-1),l,k\right)$ to the state transition probability. With these correspondences in mind we will use the terms SDTW and HMM synonymously in the following.

2.3 Model Training

The training of the symbol models works as follows: Allograph prototypes for each class are first generated by an agglomerative hierarchical clustering with the simple Euclidean DTW distance. Methods are incorporated to control the number or diversity of the clusters, resp. [1, 12].

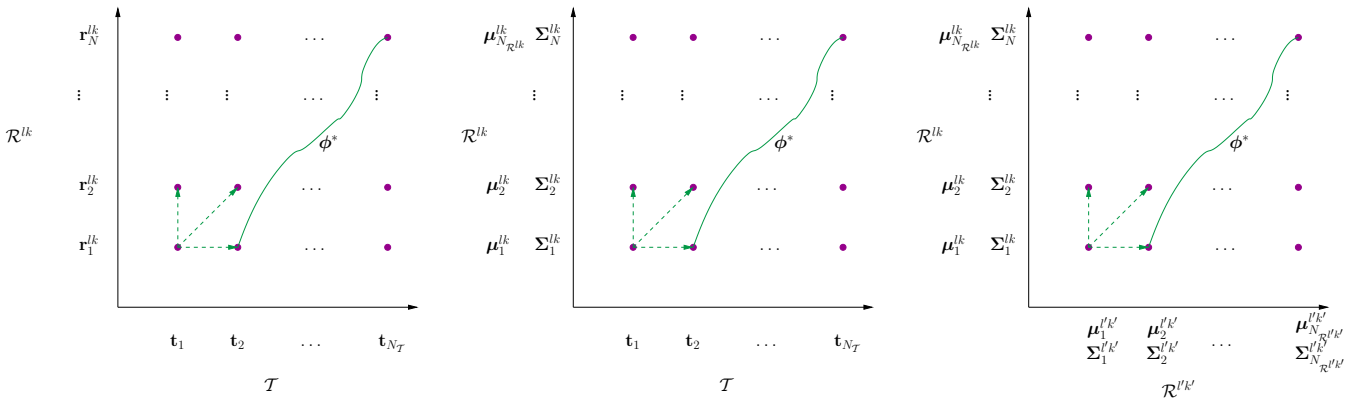
Mean, covariance and transition probabilities for each cluster (l,k) and sample point j are subsequently adapted iteratively by alternating Viterbi path detection and mean/covariance/transition-probability parameter reestimation from the samples on the Viterbi path. This training procedure is also known as Viterbi training in the HMM context. Care is needed in the estimation of the statistic of the angular quantity θ [10].

For iteration initialization the mean is set to the corresponding cluster prototypes and covariances are set to unity matrices.

3 SDTW and HMM Similarity

The framework for measuring HMM similarity is now obtained by a modification of the classifier. Looking at figure 1 (a) and (b) the difference between standard DTW and SDTW/HMM is that a pdf of the reference patterns is taken into account when computing the local distances $d(\cdot,\cdot)$.

In consequence it would therefore be straightforward to assign a pdf also to the second pattern $\mathcal{T}$ (which from now we will call $\mathcal{R}^{l'k'}$) and compute a distance $D\left(\mathcal{R}^{l'k'},\mathcal{R}^{lk}\right)$



(a) DTW **(b)** SDTW/HMM **(c)** SDTW/HMM similarity

Figure 1. Comparing the concepts of DTW, SDTW/HMM and the proposed method for computing model similarity: (a) in DTW local distances are computed from two pattern templates $\mathbf{t}_i$ and $\mathbf{r}_j^{lk}$. (b) In SDTW/HMM local distances are computed from pattern templates $\mathbf{t}_i$ and a pdf, here denoted by a second order statistic $\mathbf{R}_j^{lk} = \left(\boldsymbol{\mu}_j^{lk}, \boldsymbol{\Sigma}_j^{lk}\right)$. (c) In the proposed model similarity measure local distances are computed from two pdfs $\mathbf{R}_i^{l'k'} = \left(\boldsymbol{\mu}_i^{l'k'}, \boldsymbol{\Sigma}_i^{l'k'}\right)$ and $\mathbf{R}_j^{lk} = \left(\boldsymbol{\mu}_j^{lk}, \boldsymbol{\Sigma}_j^{lk}\right)$. Whereas local distances are defined differently in these concepts, the principle of Viterbi path detection and scoring is the same.

in a very similar way as in (2) and (3):

$$D_{\boldsymbol{\phi}}\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right) = \frac{1}{N}\sum_{i=1}^{N} d\left(\mathbf{R}^{l'k'}_{\phi_{\mathcal{R}^{l'k'}}(i)}, \mathbf{R}^{lk}_{\phi_{\mathcal{R}^{lk}}(i)}\right) \quad (5)$$

$$D\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right) = D_{\boldsymbol{\phi}^*}\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right) = \min_{\boldsymbol{\phi}}\left\{D_{\boldsymbol{\phi}}\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)\right\} \quad (6)$$

The component which has to be redefined is the local distance measure (4) of the now two pdfs. Several choices are possible, e.g. χ^2, Kullback-Leibler [5] or Jensen-Shannon divergence [8]. However, we want to interpret the model similarities in the context of experimentally gained classification errors. A measure for the probability of a classification error in a two class problem is the Bayes probability of error (or Bayes error) [12]

$$P_e\left(p_1\left(\mathbf{x}\right), p_2\left(\mathbf{x}\right)\right) = \int_{\mathbf{x}} \min\left\{\pi_1 p_1\left(\mathbf{x}\right), \pi_2 p_2\left(\mathbf{x}\right)\right\} d\mathbf{x}. \quad (7)$$

It measures the area of overlap of two pdfs p_1 and p_2 with priors π_1 and π_2 under the constraint $\pi_1 + \pi_2 = 1$ (see figure 2).

This concept of the Bayes error is now extended to the SDTW/HMM modeling. Given two models $\mathcal{R}^{l'k'}$ and $\mathcal{R}^{lk}$ and an alignment path $\boldsymbol{\phi}$, the probability of a classification

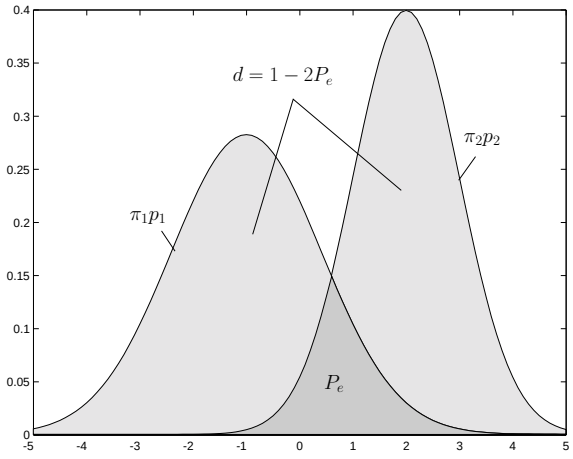


Figure 2. The Bayes probability of error $P_e\left(p_1\left(\mathbf{x}\right), p_2\left(\mathbf{x}\right)\right)$ equals the dark shaded area of overlap. The distance measure $d\left(\cdot,\cdot\right) = 1 - 2P_e\left(\cdot,\cdot\right)$ corresponds to the bright shaded area.

error in this two-class problem according to Bayes is the sum of all overlaps along the path divided by the sum of the integrals of all a-priori weighted pdfs along the path (which is N because we will normalize the priors to $\tilde{\pi}_{l'k'} + \tilde{\pi}_{lk} = 1$ and $\int_{\mathbf{x}} p(\mathbf{x})\, d\mathbf{x} = 1$), i.e.

$$P_e^{\phi}\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right) = \frac{1}{N}\sum_{i=1}^{N} P_e\left(\mathcal{N}\left(\mathbf{R}^{l'k'}_{\phi_{\mathcal{R}^{l'k'}}(i)}, \mathbf{x}\right), \mathcal{N}\left(\mathbf{R}^{lk}_{\phi_{\mathcal{R}^{lk}}(i)}, \mathbf{x}\right)\right). \tag{8}$$

From all possible alignment paths there is one which is most probable to provoke a misclassification: it is the path ϕ^* with the *maximum* overlap $P_e^{\phi^*}\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)$. In order to compute $P_e^*\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right) := P_e^{\phi^*}\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)$ and ϕ^* with our *minimization* framework, we transform the similarity measure $P_e\left(\mathcal{N}\left(\mathbf{R}_i^{l'k'}, \mathbf{x}\right), \mathcal{N}\left(\mathbf{R}_j^{lk}, \mathbf{x}\right)\right)$ into a distance measure by

$$\begin{aligned} d\left(\mathbf{R}_i^{l'k'}, \mathbf{R}_j^{lk}\right) &= 1 - 2P_e\left(\mathcal{N}\left(\mathbf{R}_i^{l'k'}, \mathbf{x}\right), \mathcal{N}\left(\mathbf{R}_j^{lk}, \mathbf{x}\right)\right) \\ &= 1 - 2\int_x \min\left\{\tilde{\pi}_{l'k'}\mathcal{N}\left(\mathbf{R}_i^{l'k'}, \mathbf{x}\right), \tilde{\pi}_{lk}\mathcal{N}\left(\mathbf{R}_j^{lk}, \mathbf{x}\right)\right\} d\mathbf{x} \end{aligned} \tag{9}$$

(i.e. the area outside the overlap, see figure 2) and apply the well known Viterbi detection (5) and (6).[1] Assuming a two-class problem the prior probabilities are normalized to $\tilde{\pi}_{l'k'} = \frac{\pi_{l'k'}}{\pi_{l'k'}+\pi_{lk}}$ and $\tilde{\pi}_{lk} = \frac{\pi_{lk}}{\pi_{l'k'}+\pi_{lk}}$.

The similarity measure $P_e^*\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)$ is achieved by back-transforming the distance $P_e^*\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right) = \frac{1}{2}\left(1 - D\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)\right)$. A nice property of $P_e^*(\cdot,\cdot)$ is that its range is $[0, 1/2]$ (and consequently $D(\cdot,\cdot) \in [0,1]$).

When integrating in (9), the periodic nature of the feature θ must be respected. This is achieved by wrapping the feature space once positively and once negatively along the periodic dimension. For that we replace $\mathcal{N}\left(\mathbf{R}_j^{lk}, \mathbf{x}\right)$ by $\mathcal{N}'\left(\mathbf{R}_j^{lk}, \mathbf{x}\right) = \sum_{i=-1}^{1}\mathcal{N}\left(\mathbf{R}_j^{lk}, \mathbf{x} + i\cdot\mathbf{s}\right)$ in the previous equations. The variable $\mathbf{s}$ is an n-dimensional vector containing a zero for non-periodic and the periodicity for periodic features ($\mathbf{s} = (0, 0, 2\pi)^T$ in our case).

We should note that the computation includes the numerical integration of the feature space, given that an analytical solution for the integration in (9) is not known in general. Consequently the complexity increases exponentially with the dimension of the feature space. For high dimension situations Monte-Carlo methods can be considered to overcome this problem. Fortunately, the applications for an HMM similarity measure mainly arise in the model training phase and thus do not require real-time computation.

[1] Minimizing $1 - 2P_e$ instead of maximizing P_e has the additional effect, that (for the chosen local transitions) the Viterbi path is not misdirected to a suboptimal solution (as indicated in section 2.2). When maximizing P_e, the Viterbi search detects a path which covers as many grid points as possible. This is due to the fact that the normalization by N is not part of the optimization criterion. Minimizing $1 - 2P_e$ gives a bias to short paths which is a more natural behavior and agrees with our Viterbi classification.

4 Results

We have trained character models with the training method described in section 2.3 and using 20% (i.e. $\approx$ 12000 samples, randomly chosen) of `1c` section (lower case characters) of the `train_r01_v07` UNIPEN online handwriting database [3]. Several clustering parameter settings which influence the diversity of clusters have been evaluated. A good compromise regarding a minimum of models and a maximum of recognition rate yielded 237 models for the 26 character classes (and 88.7% recognition rate on a disjunct test set).

In the simulation we have computed P_e^* for any of the $237^2 = 56169$ model pairs $\mathcal{R}^{lk}$ and counted the misclassifications $\mathbf{C}$ ($C_{l'l}$ is the number of samples of class l recognized as class l') for all of the $26^2 = 676$ class pairs. Both tasks were performed on the same dataset. An example of the Viterbi alignment of two reference characters "u" and "a" is shown in figure 3. For further investigation three post-processing steps have been applied for a direct comparison:

1. We reduce any set of $K_{l'} \times K_l$ allograph prototype similarities $\left\{P_e^*\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)\right\}_{k'=1,\ldots,K_{l'},k=1,\ldots,K_l}$ to one scalar value $\tilde{P}_e^*(l', l) = \sum_{k'=1}^{K_{l'}}\tilde{\pi}'_{l'k'}\sum_{k=1}^{K_l}\tilde{\pi}'_{lk}P_e^*\left(\mathcal{R}^{l'k'}, \mathcal{R}^{lk}\right)$ with $\tilde{\pi}'_{lk} = \frac{\pi_{lk}}{\sum_{i=1}^{K_l}\pi_{li}}$.

2. Since our HMM similarity assumes a two-class problem, we define the *error rate* for recognizing character l', given character l as $C'_{l'l} = C_{l'l}/\left(C_{l'l} + C_{ll}\right), \forall l' \neq l$.

3. Since $P_e^*(l', l)$ is the probability of falsely classifying class l' into l *or* l into l', we define $\tilde{C}_{l'l}$ as the error rate of classifications from class l' into l *or* l into l': $\tilde{C}_{l'l} = \tilde{\pi}''_{l'l}C'_{l'l} + \tilde{\pi}''_{ll'}C'_{ll'}$ with $\tilde{\pi}''_{l'l} = \frac{\sum_k \pi_{lk}}{\sum_k \pi_{lk} + \sum_k \pi_{l'k}}$ the prior probability of class l in the two-class context of l and l'.

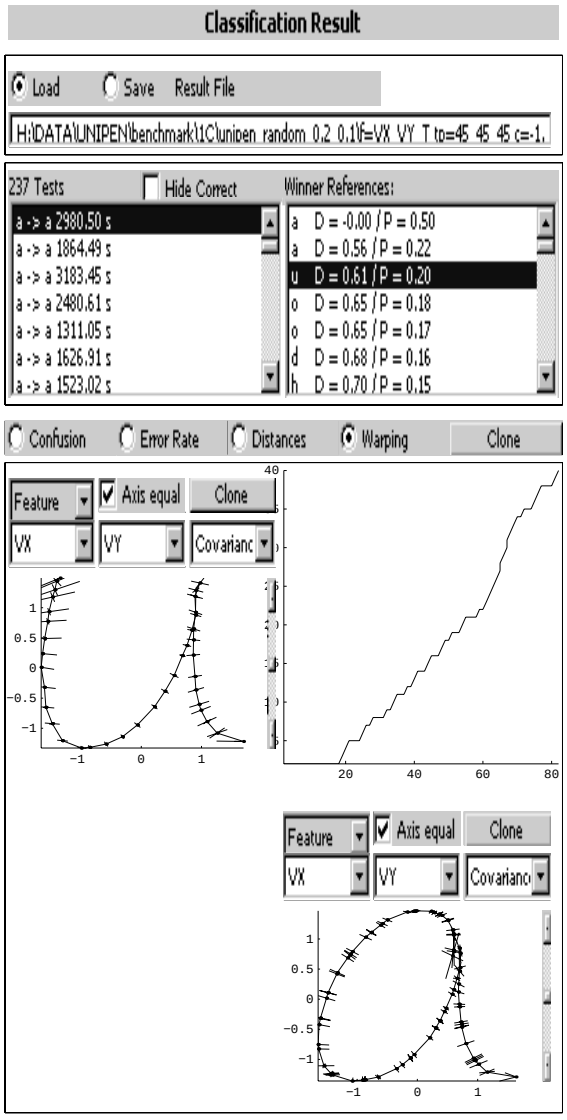
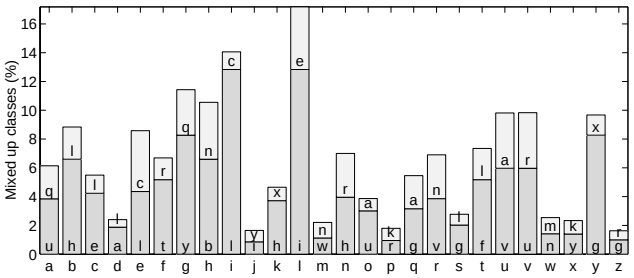
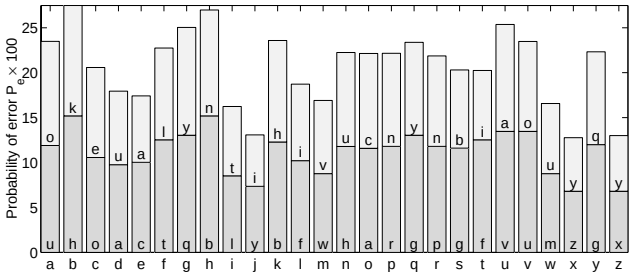


Figure 3. In the lower part of this figure two statistics "a" and "u" are illustrated with their features $\tilde{x}_i$ and $\tilde{y}_i$ with $P_e^* = 0.20$; The mean values $\boldsymbol{\mu}_j^{lk}$ are represented by small dots for each sample point, the lines at each sample point represent the projection of the first principal eigenvector/-value of $\boldsymbol{\Sigma}_j^{lk}$ in the $\tilde{x}$-$\tilde{y}$-plane by their direction and length, resp. The Viterbi alignment is shown in the third plot. In the upper part the list to the right shows the "top 7" similarity scores relative to the given "a". The first entry in the list ($D = 0$, $P_e = 0.5$) corresponds to the distance of the "a" to itself.



(a) Top 2 mixed up classes $\tilde{\mathrm{C}}$



(b) Top 2 Similarities $\tilde{P}_e^*$

Figure 4. (a) The bar diagram shows the two highest error rates $\tilde{C}_{l'l}$ for every character $l \in \{'\mathrm{a}', \ldots, '\mathrm{z}'\}$. The bars are labeled by the corresponding class names l'. (b) The stacked bars represent the similarities $\tilde{P}_e^*(l', l)$ of the two most similar classes for every character $l \in \{'\mathrm{a}', \ldots, '\mathrm{z}'\}$. The bars are labeled by the corresponding class names l'.

Results of the experiments are compared in figure 4. It is remarkable that $\tilde{P}_e^*(l', l)$ overestimates the observed error rate $\tilde{C}(l', l)$ for most of the examples shown. One reason for this might be that $\tilde{P}_e^*$ has been computed from a two-class assumption. Similarity values may have been fudged by the allograph prototype combination (see post-processing step 3). Moreover, the assumption of Gaussian pdfs may not be correct. Outliers in the data can broaden the pdfs and increase the overlap.

However, from a qualitative point of view, for 15 out of 26 classes the most similar and most frequently mixed up characters coincide (e.g. for the "a" the "u" has the highest similarity value $\tilde{P}_e^*$ and also is the most frequently mixed up character). For 25 out of 26 classes the sets of the two most similar and two most frequently mixed up classes share at least one character. It should also be noted that high values for $\tilde{P}_e^*$ also coincide with our intuitive judgment of similarity, as can be seen from figure 3 and other examples not illustrated.

The average run-time for the computation of one HMM similarity was ≈ 6 sec on a PIII 600MHz using a numerical integration resolution of $25 \times 25 \times 9$ (corresponding to $\tilde{x}, \tilde{y}, \theta$) grid points in the 3D feature space.

5 Summary and Outlook

We have presented a similarity measure for HMM based classification methods which is based on the Bayes error. The computation algorithm uses the same framework as the standard Viterbi recognition algorithm. This measure allows to analyze misclassifications, e.g. by interpreting the Viterbi state correspondences or by detecting similar model pairs. The measure can also be used as a stop criterion in the iterative HMM training or as a distance measure when clustering HMMs.

We have applied the similarity measure to the example of online handwriting character models and have shown a qualitatively close match of the most similar model pairs to misclassifications. However, the use of this measure is not limited to the application of handwriting recognition, but can also be employed in other areas where HMM based methods are used, e.g. in speech recognition or molecular biology.

It should be stressed that the suggested method is very flexible, i.e. any underlying pdf (e.g. mixed Gaussians, discrete probabilities) can be used instead of simple Gaussians. Also, the local pdf distance can be replaced by other distances like χ^2, Kullback-Leibler or Jensen-Shannon.

Interesting future work lies in the extension of this similarity measure to multi-class situations.

References

[1] D. Bockhorn. Bestimmung und Untersuchung von Signifikanzgewichtungen für die Erkennung von handgeschriebenen Buchstaben. Master's thesis, Albert-Ludwigs-Universität Freiburg, Institut für Informatik, 2000. 2.2, 2.3

[2] M. Falkhausen, H. Reininger, and D. Wolf. Calculation of distance measures between hidden Markov models. In *Proc. Eurospeech*, pages 1487–1490, 1995. 1

[3] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. UNIPEN project of on-line data exchange and recognizer benchmarks. In *Proc. of the 12th ICPR*, pages 29–33, 1994. 1, 4

[4] B. Juang and L. Rabiner. A probabilistic distance measure for hidden Markov models. *AT&T Tech. J.*, 64(2):391–408, Feb. 1985. 1

[5] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951. 3

[6] S. Kwong, Q. He, K. Man, and K. Tang. A maximum model distance approach for HMM-based speech recognition. *Pattern Recognition*, 31(3):219–229, 1998. 1

[7] S. Levinson, L. Rabiner, and M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *AT&T Tech. J.*, 62(4):1035–1074, 1983. 1

[8] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, 37(1):145–151, Jan. 1991. 3

[9] R. B. Lyngsø, C. N. S. Pedersen, and H. Nielsen. Metrics and similarity measures for hidden Markov models. In T. Lengauer et al., editors, *Proc. 7th Int. Conf. on Intelligent Syst. for Molecular Biology (ISMB-99)*, pages 178–186, 1999. 1

[10] K. Mardia. *Statistics of Directional Data*. Academic Press, London and New York, 1972. 2.3

[11] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993. 2.2, 2.2, 2.2

[12] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academix Press, 1999. 2.3, 3