IDIAP RESEARCH REPORT

USER AUTHENTICATION VIA ADAPTED GENERATIVE MODELS OF FACE IMAGES

Fabien Cardinaux [a] Conrad Sanderson [b]
Samy Bengio [c]

IDIAP–RR 04-38

JULY 2004

SUBMITTED FOR PUBLICATION

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

(a) cardinau@idiap.ch
(b) conradsand@ieee.org; Electrical and Electronic Engineering, University of Adelaide, Australia.
(c) bengio@idiap.ch

IDIAP Research Report 04-38

USER AUTHENTICATION VIA ADAPTED GENERATIVE MODELS OF FACE IMAGES

Fabien Cardinaux Conrad Sanderson Samy Bengio

JULY 2004

SUBMITTED FOR PUBLICATION

Abstract. It has been previously demonstrated that systems based on local features and relatively complex generative models, namely 1D Hidden Markov Models (HMMs) and pseudo-2D HMMs, are suitable for face recognition. Recently, a simpler generative model, namely the Gaussian Mixture Model (GMM), was also shown to perform well. In most of the previous literature related to these models, the experiments were performed with controlled images (perfect face localization, controlled lighting, background, pose, expression, etc.); however, for most secure authentication applications, the system has to be robust to more challenging conditions. In this article we evaluate the performance, robustness and complexity of GMM and HMM based approaches, using both perfect and automatic face localization, on the relatively difficult BANCA database. We also evaluate different training techniques for both GMM and HMM based systems; we show that the traditionally used Maximum Likelihood (ML) training approach has problems estimating robust model parameters when there is only a few training images available; we propose to tackle this problem through the use of Maximum *a Posteriori* (MAP) training, where the lack of data problem can be effectively circumvented. We show that models estimated with MAP are significantly more robust and are able to generalize to adverse conditions present in the BANCA database. A positive side-effect of MAP based training is that the number of client specific parameters is less than half of the number required for ML based training. We also propose to extend the GMM approach through the use of local features with embedded positional information (hence increasing performance without sacrificing the low complexity of the approach); we show that the proposed extended GMM approach obtains performance comparable to the 1D HMM approach, while being more robust and considerably less complex. We also show that while the pseudo-2D HMM approach has overall the best performance, it requires relatively long times for training and authentication.

Contents

List of Figures

List of Tables

1 Introduction

Biometric user authentication (also known as verification) has been attracting much research interest for quite some time. Applications include border control, transaction authentication, forensics and various forms of access control, such as access to digital information [1, 21, 26, 38]. In contrast to traditional access control, where approaches such as passwords and cards are used, in biometric access control a person's inherent physiological characteristics are utilized; examples of such characteristics are faces, speech, iris patterns and fingerprints.

An authentication system has to discriminate between two kinds of events: either the person claiming a given identity is the true claimant or the person is an impostor. This is related to the open set identification problem, where the task is to label a given biometric as belonging to either one of K known people or to an unknown person (i.e. a $K + 1$ classification task). An open set identification system can be implemented as an extended version of an authentication system, suggesting that research efforts should be concentrated on the authentication aspect, with later extensions being dictated by the application at hand. Both authentication and identification systems can be thought of as falling in the general research area of biometric person recognition. Further introductory and review material about the biometrics field can be found in the following papers: [11, 26, 39, 34].

The use of the face as a biometric is particularly attractive, as it can involve little or no interaction with the person to be authenticated [26]. Many techniques have been proposed for face recognition; some examples are systems based on Principal Component Analysis (PCA) feature extraction [37], modular PCA [27], Elastic Graph Matching (EGM) [9, 19], and Support Vector Machines [31], Examples specific to generative models include 1D Hidden Markov Models (HMMs) [32], pseudo-2D HMMs [12, 24] and Gaussian Mixture Models (GMMs) [4, 33, 36] (which can be considered as a simplified version of HMMs). As an in-depth review of face recognition literature is beyond the scope of this paper, the reader is directed to the following review articles: [5, 16, 18, 40].

Generative models typically use local features (that is, the features only describe a part of the face); this is in contrast to holistic features, such as in the PCA-based approach, where one feature vector describes the entire face. Local features can be obtained by analyzing a face on a block by block basis; feature extraction based on the 2D Discrete Cosine Transform (DCT) [15] or DCTmod2 [33] is usually applied to each block. In an analogous manner, 2D Gabor wavelets [20] can also be used.

In HMM based approaches, the spatial relation between major face features (such as the eyes and nose) is kept (although not rigidly); in the GMM approach the spatial relation is effectively lost (as each block is treated independently), resulting in good robustness to imperfectly located faces [4] and to out-of-plane rotations [35]. As the loss of spatial information may degrade discrimination performance, in this paper we first propose to restore some of spatial relation by using local features with embedded positional information. By working in the feature domain, the relative low-complexity advantage of the GMM approach is retained.

In the approaches presented in [12, 24, 32, 33], generative models are trained using the Maximum Likelihood (ML) criterion via the Expectation Maximization (EM) algorithm [7]. It is generally known that one of the drawbacks of training via this paradigm is that a lot of data is required to properly estimate model parameters; this can be a problem when there are only a few training images available. In an attempt to tackle this problem, Eickeler *et al.* [12] proposed to use a well trained generic (non-client specific) model as the starting point for ML training. While the results in [12] were promising, they were obtained on the rather easy Olivetti Research Ltd. (ORL) database [32]. Through experiments on the much harder BANCA database [2], we will show that even with the generic model as the starting point, ML training still produces poor models. Our second main proposition is thus to replace ML training with Maximum *a Posteriori* (MAP) training [13], which effectively circumvents the lack of data problem.

The general tone of this paper is hence an enhancement of face authentication based on generative models.

The various generative models are then compared in several image conditions (from controlled to adverse) on a common database. The comparison is done in terms of performance, robustness and complexity. We utilize perfect localization (where the eye coordinates are manually located) as well as automatic face localization.

The rest of this paper is organized as follows. In Section 2 we briefly overview the BANCA database and its experiment protocols. A summary of the automatic face localization technique is presented in Section 3. In Section 4 we overview the DCTmod2 feature extraction and describe the proposed extension (embedding of positional information). In Section 5 we review the GMM, 1D HMM and pseudo-2D HMM representations of faces; we also describe MAP training for each model. Section 6 is devoted to experiments; here we evaluate the GMM approach using standard DCTmod2 and the proposed extended features; we also evaluate the GMM, 1D and pseudo-2D HMM approaches trained with traditional ML, ML initialized by a global model, and the suggested MAP approach; lastly, we compare the complexity of the GMM and HMM based approaches. The paper is concluded in Section 7.

2 The BANCA Database and Protocols

The multi-lingual BANCA database [2] was designed to evaluate multi-modal identity authentication with various acquisition devices under several scenarios. In our experiments we use face images from the English corpus which contains 52 subjects; the population is subdivided into 2 groups of 26 subjects, denoted as $g1$ and $g2$.

Each subject participated in 12 recording sessions in different conditions and with different cameras. Each of these sessions contains two video recordings: one true client access and one impostor attack. Five "frontal" (not necessarily directly frontal) face images have been extracted from each video recording[1]. Sessions 1-4 contain data for the *controlled* condition, while sessions 5-8 and 9-12 respectively contain *degraded* and *adverse* conditions (see Fig. 1); the latter two conditions differ from the *controlled* condition in terms of image quality, lighting, background and pose.

Figure 1: Example images from the BANCA database. Left to right: *controlled*, *degraded* and *adverse* conditions.

Seven distinct configurations specify which images can be used for training and testing; the configurations are: Matched Controlled (Mc), Matched Degraded (Md), Matched Adverse (Ma), Unmatched Degraded (Ud), Unmatched Adverse (Ua), Pooled test (P) and Grand test (G). Table 1 describes the usage of different sessions in each configuration.

We believe that the most realistic cases are when we train the system in controlled conditions and test it in different conditions; hence in this paper we only performed experiments with configurations Mc, Ud, Ua and P. This limitation to four different scenarios should make the results easier to interpret.

Performance is measured in terms of Half Total Error Rate (HTER), defined as:

$$\mathrm{HTER} = (\mathrm{FAR} + \mathrm{FRR})\,/2 \tag{1}$$

[1] We note that in [17] it was shown that there are considerable improvements in authentication performance when the number of test images is increased from one to five.

Test Sessions	Train Sessions			
	1	5	9	1,5,9
C: 2-4 I: 1-4	Mc			
C: 6-8 I: 5-8	Ud	Md		
C: 10-12 I: 9-12	Ua		Ma	
C: 2-4,6-8,10-12 I: 1-12	P			G

Table 1: Usage of the seven BANCA protocols (C: client, I: impostor)

where FAR and FRR are the False Acceptance Rate and False Rejection Rate, respectively. The HTER is a particular case of the Decision Cost Function (DCF) [3, 8]:

$$\text{DCF} = \text{Cost(FR)} \cdot P(\text{client}) \cdot \text{FRR} + \text{Cost(FA)} \cdot P(\text{impostor}) \cdot \text{FAR} \tag{2}$$

where $P(\text{client})$ is the prior probability that a client will use the system, $P(\text{impostor})$ is the prior probability that an impostor will use the system, Cost(FR) is the cost of a false rejection and Cost(FA) is the cost of a false acceptance. For the HTER, we have $P(\text{client})$=$P(\text{impostor})$=0.5 and the costs are set to 1.

Since in real life the decision threshold has to be chosen *a priori*, it is selected to obtain Equal Error Rate (EER) performance (where FAR=FRR) on the validation set; it is then used on the test set to obtain a HTER figure. Here we use set $g1$ as the validation set and set $g2$ as the test set.

3 Face Localization

Face recognition results in the literature are usually presented assuming perfect localization (e.g. see [12, 24, 25, 32, 33, 36]); in only relatively few publications we find performance evaluation using an automatic face localization system (e.g. [4, 31]). While assuming perfect localization makes the results independent of the quality of the face localization system, they are biased when compared to a real life system, where the face needs to be automatically located. In this study we present results for both cases.

For "perfect localization" experiments, we use the manually annotated eye center positions, as provided with the BANCA database [2]. For "automatic localization" experiments, we use a face locator based on a multi-scale scanning approach, similar in nature to the system presented in [14]. Briefly, each given image is downsampled to produce several scaled versions; each scaled version is exhaustively scanned using a 32×32 pixel window; on each window, holistic PCA-based feature extraction is performed, with the result being a 128 dimensional feature vector. A GMM based classifier (see Section 5) is used to find the likelihood of the feature vector representing a face as well as the the likelihood of representing a non-face. The difference between the two likelihoods (see Eqn. (3)) is used as a "probability of face" measure; each location in each scaled image is given a corresponding "probability of face" measure; the location of the face (and the implied size of the face) is then found by finding the location of the highest measure. Specific to the face localization task, the PCA-based feature extraction as well as the GMM based classifier were trained using images from the FERET [28], XM2VTS [23] and VidTIMIT [33] databases; for the non-face class, the GMM based classifier used training images of scenes not containing faces (e.g. forests, mountains, city scapes, buildings, rooms, etc).

4 Preprocessing and Feature Extraction

Face images from the BANCA database are converted into gray-scale values and a 80×64 (rows $\times$ columns) face window is cropped out; face location is either based on manually located eye positions or using the automatic face localization system. Each face window contains the face area from the eyebrows to the mouth; moreover, the location of the eyes is the same for each face window (via geometric normalization); see Fig. 2 for an example.

Histogram equalization is used to normalize the face images photometrically. We then extract *DCTmod2* features from each image face [33]. We have found the combination of histogram equalization and feature extraction to provide good results in preliminary experiments. The feature extraction process is summarized as follows. A given face image is analyzed on a block by block basis; each block has a size of $N_P \times N_P$ pixels (here we use $N_P = 8$) and overlaps neighboring blocks by a configurable amount of pixels. Each block is decomposed in terms of two-dimensional DCT basis functions [15]. A feature vector for a block located at row a and column b is then constructed as:

$$\vec{x}_{(a,b)} = \left[\; \Delta^h c_0 \; \Delta^v c_0 \; \Delta^h c_1 \; \Delta^v c_1 \; \Delta^h c_2 \; \Delta^v c_2 \;\; c_3 \; c_4 \; ... \; c_{M-1} \; \right]^T$$

where c_n represents the n-th DCT coefficient, while $\Delta^h c_n$ and $\Delta^v c_n$ represent the horizontal and vertical delta coefficients respectively, and are computed using DCT coefficients extracted from neighboring blocks. Compared to traditional DCT feature extraction [12, 24], the first three DCT coefficients are replaced by their respective horizontal and vertical deltas in order to reduce the effects of illumination direction changes. In this study we use M=15 (choice based on [33]), resulting in an 18 dimensional feature vector for each block.

The degree of overlap has two effects: the first is that as overlap is increased the spatial area used to derive one feature vector is decreased; the second is that as the overlap is increased the number of feature vectors extracted from an image grows in a quadratic manner.

4.1 Embedding Positional Information

The above DCTmod2 feature extraction has been successfully used in GMM based face authentication systems [4, 33]. In such GMM systems (see Section 5.1 for more details) the spatial relation between major face features (such as the eyes and nose) is effectively lost (as each block is treated independently). As the spatial relations can provide discriminatory information, we propose to increase the performance of the GMM approach (without sacrificing its simplicity) through extending the DCTmod2 approach with embedded positional information. Formally, the feature vector for a block at row a and column b is found with:

$$\vec{x}_{(a,b)}^{\,\texttt{extended}} = \quad \vec{x}_{(a,b)}^{\,\texttt{original}} \;\; {}^T \;\; a \;\; b \;\;\; {}^T$$

where $\vec{x}_{(a,b)}^{\,\texttt{original}}$ is the standard DCTmod2 feature vector for the block located at row a and column b. By explicitly embedding positional information into each feature vector, a weak constraint is placed on the areas that each gaussian in the GMM can model, thus making a face model more specific.

5 Generative Model Based Classifiers

Let us denote the parameter set for client C as λ_C, and the parameter set describing a generic face (non-client specific) as $\lambda_{\overline{C}}$. Given a claim for client C's identity and a set of feature vectors $X = \{\vec{x}_t\}_{t=1}^{N_V}$ supporting the claim (extracted from the given face), we find an opinion on the claim using:

$$\Lambda(X) = \log P(X|\lambda_C) - \log P(X|\lambda_{\overline{C}}) \tag{3}$$

where $P(X|\lambda_C)$ is the likelihood of the claim coming from the true claimant and $P(X|\lambda_{\overline{C}})$ is the likelihood of the claim coming from an impostor. The generic face model is also known as a *World Model* and a *Universal*

Background Model [22, 30]; it is typically trained with data from many people (here we use data from BANCA's Spanish corpus). The authentication decision is then reached as follows: given a threshold τ, the claim is accepted when $\Lambda(X) \geq \tau$ and rejected when $\Lambda(X) < \tau$.

We use three different ways to train each client model:

1. Traditional ML training, where k-means initialization is used [7, 10].

2. ML training with a generic (non-client specific) model as the starting point (as in [12]); data from many people is used the find the parameters of the generic model via traditional ML training; this is the same generic model used for calculating $P(X|\lambda_{\overline{C}})$ in Eqn. (3) for all generative approaches.

3. MAP training [13]; here a generic model is used as in point (2) above, but instead of using it merely as a starting point, the model is *adapted* using client data. Given a set of training vectors, X, the probability density function (pdf) $P(X|\lambda)$ and the prior pdf of λ, $P(\lambda)$, the MAP estimate of model parameters, $\lambda_{\texttt{MAP}}$, is defined as:

$$\lambda_{\texttt{MAP}} = \arg\max_{\lambda} P(\lambda|X) \tag{4}$$

$$= \arg\max_{\lambda} P(X|\lambda)P(\lambda) \tag{5}$$

Assuming λ to be fixed but unknown is equivalent to having a non-informative $P(\lambda)$, reducing the solution of $\lambda_{\texttt{MAP}}$ to the standard ML solution. Thus, the difference between ML and MAP training is in the definition of the prior distribution for the model parameters to be estimated. Further discussion on MAP training is given in Section 5.1.

5.1 Gaussian Mixture Model

In the GMM approach, the likelihood of a set of feature vectors is found with

$$P(X|\lambda) = \prod_{t=1}^{N_V} P(\vec{x}_t|\lambda) \tag{6}$$

where

$$P(\vec{x}|\lambda) = \sum_{j=1}^{N_G} m_j \, \mathcal{N}(\vec{x}|\vec{\mu}_j, \mathbf{\Sigma}_j) \tag{7}$$

$$\lambda = \{m_j, \vec{\mu}_j, \mathbf{\Sigma}_j\}_{j=1}^{N_G} \tag{8}$$

Here, $\mathcal{N}(\vec{x}|\vec{\mu}, \mathbf{\Sigma})$ is a D-dimensional gaussian density function with mean $\vec{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$. N_G is the number of gaussians and m_j is the weight for gaussian j (with constraints $\sum_{j=1}^{N_G} m_j = 1$ and $\forall\, j : m_j \geq 0$).

An implementation of MAP for client model adaptation consists of using a global parameter to tune the relative importance of the prior. In this case, the equation for adaptation of the means is [22]:

$$\hat{\mu}_k = (1-\alpha)\mu_k^w + \alpha \frac{\sum_{t=1}^{T} P(m_t = k|\vec{x}_t)\vec{x}_t}{\sum_{t=1}^{T} P(m_t = k|\vec{x}_t)} \tag{9}$$

where $\hat{\mu}_k$ is the new mean of the k-th gaussian, μ_k^w is the corresponding mean in the generic model, $P(m_t = k|\vec{x}_t)$ is the posterior probability of the k-th gaussian (from the client model from the previous iteration), and $\alpha \in [0, 1]$ is the adaptation factor chosen empirically on a separate validation set. The adaptation procedure is iterative, thus an initial client model is required; this is accomplished by copying the generic model.

As can be seen, the new mean is simply a weighted sum of the prior mean and new statistics; α can hence be interpreted as the amount of faith we have in the new statistics; when the amount of training data is low, we would generally set α to be low.

It must be noted that only the means of the gaussians are adapted, as it has been empirically observed that adaptation of the other parameters generally does not improve performance [22]. The other parameters (the weights and covariance matrices) are copied from the generic model to each client model.

5.2 1D Hidden Markov Model

The one-dimensional HMM (1D HMM) is a particular HMM topology where only self transitions or transitions to the next state are allowed. This type of HMM is also known as a top-bottom HMM [32] or left-right HMM in the context of speech recognition [29]. Here the face is represented as a sequence of overlapping *rectangular* blocks from top to bottom of the face (see Fig. 2 for an example). To simulate the *rectangular* block representation, DCTmod2 feature vectors from the same row of blocks are concatenated to form a large observation vector.

The model is characterized by the following:

1. N, the number of states in the model; each state corresponds to a region of the face; $S = \{S_1, S_2, \ldots, S_N\}$ is the set of states. The state of the model at row t is given by $q_t \in S$, $1 \leq t \leq T$, where T is the length of the observation sequence (number of rectangular blocks).

2. The state transition matrix $A = \{a_{ij}\}$. The topology of the 1D HMM allows only self transitions or transitions to the next state:

$$a_{ij} = \begin{cases} P(q_t = S_j | q_{t-1} = S_i) & \text{for } j = i,\ j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

3. The state probability distribution $B = \{b_j(\vec{x}_t)\}$, where

$$b_j(\vec{x}_t) = p(\vec{x}_t | q_t = S_j) \tag{10}$$

The features are expected to follow a continuous distribution and are modeled with mixtures of gaussians.

In compact notation, the parameter set of the 1D HMM is:

$$\lambda = (A, B) \tag{11}$$

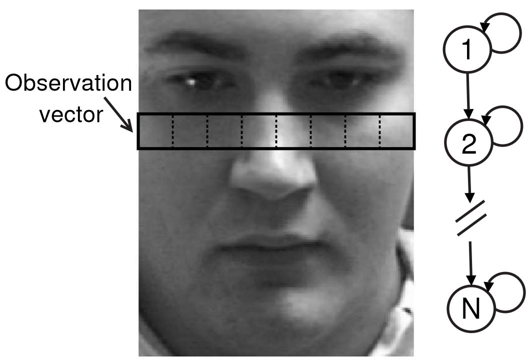

Figure 2: Sampling window and 1D HMM topology.

If we let Q to be a state sequence $q_1, q_2 \cdots q_T$, then the likelihood of an observation sequence X is:

$$P(X|\lambda) = \sum_{\forall\, Q} P(X, Q|\lambda) \tag{12}$$

$$= \sum_{\forall\, Q} \prod_{t=1}^{T} b_{q_t}(\vec{x}_t) \prod_{t=2}^{T} a_{q_{t-1},q_t} \tag{13}$$

The calculation of this likelihood according to the direct definition (13) involves an exponential number of computations; in practice the Forward-Backward procedure is used [29]; it is mathematically equivalent, but significantly more efficient.

For the case of the 1D HMM, MAP adaptation of the means is [c.f. Eqn. (9)]:

$$\hat{\mu}_{k,i} = (1-\alpha)\mu^w_{k,i} + \alpha \frac{\sum_{t=1}^{T} P(q_t = i|\vec{x}_t) P(m^i_t = k|\vec{x}_t)\vec{x}_t}{\sum_{t=1}^{T} P(q_t = i|\vec{x}_t) P(m^i_t = k|\vec{x}_t)} \tag{14}$$

where $P(q_t = i|\vec{x}_t)$ is the posterior probability of the state i at row t and $P(m^i_t = k|\vec{x}_t)$ is the posterior probability of its k-th gaussian.

5.3 Pseudo-2D HMM

Emission probabilities of 1D HMMs are typically represented using mixtures of gaussians. For the case of P2D HMM, the emission probabilities of the HMM (now referred to as "main HMM") are estimated through a secondary HMM (referred to as an "embedded HMM"). The states of the embedded HMMs are in turn modeled by a mixture of gaussians. This approach was used for the face identification task in [12, 32] and the training process is described in detail in [25]. As shown in Fig. 3, we chose to perform the vertical segmentation of the face image by the main HMM and horizontal segmentation by embedded HMMs. We made this choice because the main decomposition of the face is instinctively from top to bottom (forehead, eyes, nose, mouth). Note that the opposite choice has been made in [12, 32]. It is important to note that the segmentation using this HMM topology constrains the segmentation done by the main HMM to be the same for all columns (if the main HMM performs the vertical segmentation) or all rows (if the main HMM performs the horizontal segmentation). The

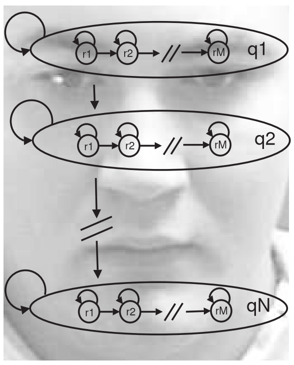

Figure 3: P2D HMM: the emission distributions of the vertical HMM are estimated by horizontal HMMs. q_i represent the states of the main HMM and r_j represent the embedded HMMs states.

corresponding equation for MAP adaptation of the means [c.f. Eqns. (9) and (14)] is:

$$\hat{\mu}_{k,i,j} = (1-\alpha)\mu^{w}_{k,i,j} + \alpha\hat{\mu}^{ML}_{k,i,j} \tag{15}$$

with:

$$\hat{\mu}^{ML}_{k,i,j} = \frac{\sum_{t=1}^{T} P(q_t = i|\vec{x}_t)P(r^i_t = j|\vec{x}_t)P(m^{i,j}_t = k|\vec{x}_t)\vec{x}_t}{\sum_{t=1}^{T} P(q_t = i|\vec{x}_t)P(r^i_t = j|\vec{x}_t)P(m^{i,j}_t = k|\vec{x}_t)} \tag{16}$$

where $P(q_t = i|\vec{x}_t)$ is the posterior of the state i of the main HMM, $P(r^i_t = j|\vec{x}_t)$ is the posterior of the state j of its embedded HMM and $P(m^{i,j}_t = k|\vec{x}_t)$ is the posterior of its k-th gaussian.

6 Experiments and Discussion

For each client model, the training set was composed of five images extracted from the same video sequence. We artificially increased this to ten images by mirroring each original face window. The generic model was trained with 571 face images (extended to 1142 by mirroring) from the Spanish corpus of BANCA (containing faces different from the English corpus), thus making the generic model independent of the subjects present in the client database. DCTmod2 features were extracted using either a four or a seven pixel overlap; using the validation set $g1$ we found that an overlap of four pixels is better for the GMM approaches while an overlap of seven pixels is better for the HMM based approaches. The effects of the differences in the overlap are currently under further investigation.

As described in Section 5.2, feature vectors from the same row are concatenated in the 1D HMM approach. Since the resulting vector would be too big if we concatenate all the features from the same row (recall that seven pixel overlap is used), we chose to concatenate features from every eighth block (thus eliminating horizontally overlapped blocks); this resulted in 126 dimensional feature vectors for each rectangular block.

In order to optimize each model, we used the validation set $g1$ to select the size of the model (e.g. number of gaussians in the GMMs and number of states of the HMMs) as well as other hyper-parameters, such as the adaptation coefficient α, the threshold τ and the variance floor (lower bound on variance parameters) for the related generic model. The hyper-parameters were chosen to minimize the EER. The performance of each model was then tested on set $g2$.

After optimization, the GMM based approaches used 512 gaussians per client, while the 1D HMM approach used 10 states and 12 gaussians per state (for a total of 120 gaussians) and the P2D HMM approach used 10 states in the main HMM, 4 states in each embedded HMM and 20 gaussians per state (for a total of 800 gaussians). It must be noted that the number of gaussians is not a complete description of the complexity of each model (see Section 6.3).

6.1 Perfect Localization

Table 2 shows the results in terms of HTER for a perfect localization and for the four different systems presented in this paper.

Specifically, GMM indicates the GMM approach with standard DCTmod2 feature vectors, GMMext indicates the GMM approach with extended DCTmod2 feature vectors, and 1D HMM & P2D HMM are self explanatory. For all four systems results are shown for the three different training strategies (Section 5); models trained using the traditional ML criterion have a *ML* suffix; for ML training initialized with a generic model, the suffix is *init*; for MAP training, the suffix is *adapt*.

For comparison purposes, the table also contains performance figures for three baseline holistic systems, namely a PCA based system, a system based on combination of Linear Discriminant Analysis and Normalized Correlation (LDA/NC), and a system based on Support Vector Machines (SVMs); the latter two systems are the two best systems reported in [31]. In contrast to the approaches described in Section 5 (where each feature

System	Protocol			
	Mc	Ud	Ua	P
PCA	5.1	26.1	22.0	21.6
LDA/NC (from [31])	4.9	16.0	20.2	$*$ 14.8
SVM (from [31])	5.4	25.4	30.1	20.3
GMM *ML*	**5.5**	44.6	26.0	26.6
GMM *init*	**5.5**	45.0	25.8	26.5
GMM *adapt*	6.4	**25.6**	**22.8**	**19.4**
GMMext *ML*	5.6	38.8	21.3	23.9
GMMext *init*	**5.1**	37.2	21.2	23.8
GMMext *adapt*	6.2	**23.7**	**17.6**	**18.6**
1D HMM *ML*	$*$ **2.4**	26.6	21.8	21.6
1D HMM *init*	5.1	27.4	21.8	21.9
1D HMM *adapt*	6.9	**16.0**	**17.3**	**19.8**
P2D HMM *ML*	8.3	27.0	23.0	22.1
P2D HMM *init*	10.1	25.5	22.6	22.0
P2D HMM *adapt*	**3.4**	$*$ **12.7**	$*$ **15.4**	**16.4**

Table 2: HTER performance for **perfect localization** using GMM (standard DCTmod2 features), GMMext (extended DCTmod2 features), 1D HMM and P2D HMM. **ML**: client models trained using traditional ML criterion; **init**: client models trained using ML initialized with a generic model; **adapt**: client models trained using MAP. The asterix indicates the best result for a protocol, while boldface indicates the best result within a model type and protocol.

vector represents only a part of the face), a holistic representation takes into account the *entire* face when deriving *one* feature vector (in other words, the spatial relation between face characteristics is rigidly kept).

The results for systems based on LDA/NC and SVMs are taken directly from [31], while the PCA based system is our own implementation. It must be noted that in [31], $g1$ and $g2$ were used alternatively as the validation set and the test set; the results were then computed using the mean of HTERs from the two configurations; in contrast we have performed our experiments only with the $g1$ as the validation set and $g2$ as the test set.

For the PCA based system we utilized 160-dimensional feature vectors (one vector per face); the dimensionality was found to provide optimal performance in preliminary experiments. The classifier used for the PCA system is somewhat similar to the local feature GMM approach. The main difference is that only two gaussians are utilized: one for the client and one to represent the generic model. Due to the very small amount of training data, each client model inherits the covariance matrix from the generic model and the mean of each client model is the mean of the training vectors for that client. A similar system has been used in [35, 34].

It is interesting to see that for the Matched Controlled condition (Mc), ML training performs better than adaptation (except for P2D HMM as discussed later) but for Unmatched conditions (Ud and Ua) or partially unmatched condition (P) the models trained by MAP always perform better. We believe that the models trained by ML are too highly tuned (i.e. over-fitted) to the training data (and hence the training condition); while this works well when the training and testing conditions are matched (as in Mc), when the condition changes there is a mismatch between the model and the given condition, resulting in rapid performance degradation. The results also show that ML training with initialization by a generic model generally does not eventuate in better models compared to traditional ML training (where k-means initialization is used).

For the GMM approach, we can see that the use of extended DCTmod2 feature vectors results in better performance compared to standard DCTmod2, especially in the Ua condition. It can also be seen that the performance of the extended GMM approach is comparable to the 1D HMM approach.

In the 1D HMM approach, the dimensionality of the feature vectors is 126 against 18 for the standard GMM and P2D HMM approaches and 20 for the extended GMM approach. We believe that the large dimensionality of the feature vectors used in the 1D HMM approach is the main drawback; the larger the dimensionality, the more training data is required to properly estimate model parameters [10] (especially for the generic model, which is then adapted for each client).

When using ML training, the performance of P2D HMM approach is not better than 1D HMM; this can be explained by the much larger number of parameters used in P2D HMM (hence requiring more training data). However, when MAP training is used, the lack of data problem is effectively circumvented, resulting in the P2D HMM approach obtaining (in almost all cases) significantly better performances than the other generative models presented in this paper; moreover, in three out of four cases, the P2D HMM system performs better than the LDA/NC system presented in [31].

6.2 Automatic Localization

Table 3 shows the results in terms of HTER when using an automatic face localization system. For this set of experiments we used exactly the same models (trained with perfectly localized faces) as for the previous experiments. It must be noted that the LDA/NC and SVM based systems in [31] utilize a different face localization system from the one used for the PCA, GMM and HMM systems; as such the results are not directly comparable, but are included as an example of the performance degradation that occurs when automatic face localization is utilized (compared to using perfectly located faces).

System	Protocol			
	Mc	Ud	Ua	P
PCA	25.2	37.2	33.0	32.9
LDA/NC (from [31])	22.6	25.4	27.1	25.2
SVM (from [31])	19.7	30.4	33.2	27.8
GMM *ML*	15.9	44.4	30.3	30.7
GMM *init*	15.2	44.9	28.7	30.8
GMM *adapt*	∗ **8.8**	**30.9**	**26.3**	**22.9**
GMMext *ML*	13.3	41.3	26.1	27.2
GMMext *init*	13.8	41.2	26.9	28.4
GMMext *adapt*	**12.2**	**24.7**	**21.5**	**21.7**
1D HMM *ML*	**13.0**	28.8	**19.7**	25.0
1D HMM *init*	14.3	36.2	22.9	29.0
1D HMM *adapt*	16.7	**27.2**	20.0	**24.9**
P2D HMM *ML*	14.1	28.5	24.4	22.7
P2D HMM *init*	13.1	26.4	25.0	22.9
P2D HMM *adapt*	**9.6**	∗**18.9**	∗**17.5**	∗**18.4**

Table 3: As per Table 2 but using **automatic face localization**.

It can be observed that the results are consistent with results for perfect localization, i.e. the adaptation training approach is the most appropriate in almost all the cases. When looking only at models trained via the adaptation approach, the P2D HMM based system provides the best performance, followed by GMMext, 1D HMM and finally GMM. By comparing Tables 2 and 3 it can be deduced that (when using adaptation based training) the most robust system (i.e. with the least difference in performance between perfect and automatic face localization) is GMMext, followed by GMM, P2D HMM and finally 1D HMM. Furthermore, we can see the PCA based system is the most affected by imperfect face localization. We conjecture that the poor performance of the PCA based system is due to the rigid preservation of spatial relations between face characteristics; this is in contrast to local feature based approaches, where the spatial relation between facial characteristics is inherently less constrained; hence translations due to imperfect face localization have less impact on local feature based approaches.

6.3 Complexity of Models

Apart from the performance, the complexity of a given model is also an important consideration; here, by “complexity” we mean the number of parameters to store for each client as well as the time required for training and authentication. If we wish to store each model on an electronic card (e.g. a credit card), the size of the model becomes an important issue. We are specifically interested in the number of *client specific* parameters, meaning that we count only parameters which are different between the clients.

Table 4 shows the complexity of each local feature model used in our experiments (using optimal hyper-parameters, such as the number of gaussians); specifically, we show the number of client specific parameters, the time taken to train the world model, the client model training time, and the time required to authenticate one claim (comprised of five images). The experiments were done on a Pentium IV 3 GHz running Red Hat Linux 7.3. The times include pre-processing time; the values in brackets indicate the time for authentication or training excluding steps such as face localization, normalization and feature extraction. While the implementation of GMM and HMM based systems was not specifically optimized in terms of computation time, we believe the times presented are indicative.

Model type	GMM			GMMext			1D HMM			P2D HMM		
Training type	ML	init	adapt	ML	init	adapt	ML	init	adapt	ML	init	adapt
number of client specific parameters	18,944		9,216	20,992		10,240	30,379		15,120	29,689		14,400
world model training time	525s (392s)			728s (595s)			988s (810s)			32,234s (32,056s)		
client model training time	18s (16s)	19s (16s)	9s (8s)	18s (16s)	19s (18s)	9s (8s)	20s (19s)	20s (19s)	11s (10s)	923s (921s)	812s (811s)	359s (358s)
time for authentication of one claim (5 images)	1.06s (0.18s)			1.06s (0.18s)			1.36s (0.27s)			6.74s (5.65s)		

Table 4: **Complexity of the models**. Times are given in terms of seconds. Values in brackets exclude pre-processing time (e.g. face localization, normalization, feature extraction, etc).

The number of client specific parameters for GMM based approaches is the sum of the parameters for the means, covariance matrices (both dependent on the dimensionality of feature vectors) and weights; for the HMM based approaches transition probabilities are also taken into account. When MAP training is used, only the means need to be counted, since the other parameters are shared by all clients; the shared parameters can be stored only once in the system for all clients (e.g. there is no need to store them in each client’s electronic card). This is in contrast to ML based training, where there are no parameters shared between client models. To put it another way, when utilizing MAP based training the number of client specific parameters is less than half of the number required for ML based training.

We can see that the HMM based approaches have considerably more client specific parameters that the GMM based approaches. Taking into account the results presented in Table 3, it can be observed that even though the 1D HMM approach utilizes more parameters than the GMMext approach, its overall performance, for automatically located faces and MAP training, is actually worse.

Training of the generic model can be done off-line and hence the time required is not of great importance; however, the time taken to train each client model as well as the time for one authentication are quite important. There shouldn’t be a long delay between a user enrolling in the system and being able to use the system;

most importantly, the authentication time should not be cumbersome, in order to aid the adoption of the authentication system. When utilizing MAP training, the training time is generally halved when compared to ML based training. The GMM, GMMext and 1D HMM approaches have a reasonable authentication time of around one second. The P2D HMM approach has a considerably higher authentication time, at around six seconds; with current computing resources, this authentication time can be considered as being too long for practical deployment purposes.

7 Conclusions and Future Work

It has been previously demonstrated that systems based on local features and relatively complex generative models, namely 1D Hidden Markov Models (HMMs) and pseudo-2D HMMs, are suitable for face recognition. Recently, a simpler generative model, namely the Gaussian Mixture Model (GMM), was also shown to perform well. In most of the previous literature related to these models, the experiments were performed with controlled images (perfect face localization, controlled lighting, background, pose, expression, etc.); however, for most secure authentication applications, the system has to be robust to more challenging conditions.

In this article we evaluated the performance, robustness and complexity of GMM and HMM based approaches, using both perfect and automatic face localization, on the relatively difficult BANCA database. We also evaluated different training techniques for both GMM and HMM based systems; we showed that the traditionally used Maximum Likelihood (ML) training approach has problems estimating robust model parameters when there is only a few training images available; we proposed to tackle this problem through the use of Maximum *a Posteriori* (MAP) training, where the lack of data problem can be effectively circumvented. We showed that models estimated with MAP are significantly more robust and are able to generalize to adverse conditions present in the BANCA database. A positive side-effect of MAP based training is that the number of client specific parameters is less than half of the number required for ML based training.

We also proposed to extend the GMM approach through the use of local features with embedded positional information (hence increasing performance without sacrificing the low complexity of the approach); we showed that the proposed extended GMM approach obtains performance comparable to the 1D HMM approach, while being more robust and considerably less complex. We also showed that while the pseudo-2D HMM approach has overall the best performance, it requires relatively long times for training and authentication. For the GMM and the 1D HMM based approaches, the preprocessing severely penalizes the speed of the authentication; efforts should thus be made on this part to speed up the overall system.

In future work we will investigate effects of embedding positional information into feature vectors used in the P2D HMM approach; moreover, we will examine why different overlap settings in DCTmod2 feature extraction are preferred by different models.

Acknowledgments

The authors thank the Swiss National Science Foundation for supporting this work through the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2). The GMM and HMM systems were implemented with the aid of the *Torch* machine learning library [6].

References

[1] W. Atkins, “A testing time for face recognition technology”, *Biometric Technology Today*, Vol. 9, No. 3, 2001, pp. 8-11.

[2] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, J.-P. Thiran, “The BANCA Database and Evaluation Protocol”, *Proc. Audio- and Video-Based Biometric Person Authentication*, Guildford, 2003, pp. 625-638.

[3] S. Bengio, J. Mariéthoz, S. Marcel, “Evaluation of Biometric Technology on XM2VTS”, IDIAP Research Report 01-21, Martigny, Switzerland, 2001.

[4] F. Cardinaux, C. Sanderson, S. Marcel, “Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS”, *Proc. Audio- and Video-Based Biometric Person Authentication*, Guildford, 2003, pp. 911-920.

[5] R. Chellappa, C. L. Wilson, S. Sirohey, “Human and Machine Recognition of Faces: A Survey”, *Proceedings of the IEEE*, Vol. 83, No. 5, 1995, pp. 705-740.

[6] R. Collobert, S. Bengio, J. Mariéthoz, “Torch: a modular machine learning software library”, IDIAP Research Report 02-46, Martigny, Switzerland, 2002.

[7] A.P. Dempster, N.M. Laird, D.B. Rubin, “Maximum-likelihood from incomplete data via the EM algorithm”, *J. Royal Statistical Soc. Ser. B*, Vol. 39, No. 1, 1977, pp. 1-38.

[8] G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds, “The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective”, *Speech Communication*, Vol. 31, No. 2-3, 2000, pp. 225-254.

[9] B. Duc, S. Fischer, J. Bigün, “Face Authentication with Gabor Information on Deformable Graphs”, *IEEE Trans. Image Processing*, Vol. 8, No. 4, 1999, pp. 504-516.

[10] R. Duda, P. Hart, G. Stork, *Pattern Classification*, Wiley, 2001.

[11] J.-L. Dugelay, J.-C. Junqua, C. Kotropoulos, R. Kuhn, F. Perronnin, I. Pitas, “Recent Advances in Biometric Person Authentication”, *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Orlando, 2002, Vol. IV, pp. 4060-4062.

[12] S. Eickeler, S. Müller, R. Gerhard, “Recognition of JPEG Compressed Face Images Based on Statistical Methods”, *Image and Vision Computing*, Vol. 18, No. 4, 2000, pp. 279-287.

[13] J.-L. Gauvain and C.-H. Lee, “Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, 1994, pp. 291-298.

[14] L. Chengjun, “A Bayesian discriminating features method for face detection”, *Pattern Analysis and Machine Intelligence*, Vol. 25, No. 6, 2003, pp. 725-740.

[15] R. C. Gonzales and R. E. Woods, *Digital Image Processing*, Addison-Wesley, 1992.

[16] M. A. Grudin, “On internal representations in face recognition systems”, *Pattern Recognition*, Vol. 33, No. 7, 2000, pp. 1161-1177.

[17] J. Kittler, J. Matas, K. Johnsson, M.U. Ramos-Sánchez, “Combining evidence in personal identity verification systems”, *Pattern Recognition Letters*, Vol. 18, No. 9, 1997, pp. 845-852.

[18] S.G. Kong, J. Heo, B.R. Abidi, J. Paik, M.A. Abidi, "Recent advances in visual and infrared face recognition - a review", *Computer Vision and Image Understanding*, in press.

[19] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R.P. Würtz, W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture", *IEEE Trans. Computers*, Vol. 42, No. 3, 1993, pp. 300-311.

[20] T. S. Lee, "Image Representation Using 2D Gabor Wavelets", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 18, No. 10, 1996, pp. 959-971.

[21] M. Lockie (editor), "Facial verification bureau launched by police IT group", *Biometric Technology Today*, Vol. 10, No. 3, 2002, pp. 3-4.

[22] J. Mariéthoz and S. Bengio, "A Comparative Study of Adaptation Methods for Speaker Verification", *Proc. Int. Conf. Spoken Language Processing*, Denver, 2002, pp. 581-584.

[23] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database", *Proc. Int. Conf. Audio and Video-based Biometric Person Authentication*, Washington D.C., 1999, pp. 72-77.

[24] A. Nefian and M. Hayes, "Face recognition using an embedded HMM", *Proc. Audio and Video-based Biometric Person Authentication*, Washington D.C., 1999, pp. 19-24.

[25] A. Nefian and M. Hayes, "Maximum likelihood training of the embedded HMM for face detection and recognition", *Proc. IEEE Int. Conf. Image Processing*, Vancouver, 2000, Vol. 1, pp. 33-36.

[26] J. Ortega-Garcia, J. Bigün, D. Reynolds, J. Gonzales-Rodriguez, "Authentication Gets Personal with Biometrics", *IEEE Signal Processing Magazine*, Vol. 21, No. 2, 2004, pp. 50-62.

[27] A. Pentland, B. Moghaddam, T. Starner, "View-Based and Modular Eigenspaces for Face Recognition", *Proc. Int. Conf. Computer Vision and Pattern Recognition*, Seattle, 1994, pp. 84-91.

[28] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10, 2000, pp. 1090-1104.

[29] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", in: *Readings in Speech Recognition* (eds.: A. Waibel and K.-F. Lee), Kaufmann, San Mateo, 1990, pp. 267-296.

[30] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, No. 1-3, 2000.

[31] M. Sadeghi, J. Kittler, A. Kostin, K. Messer, "A Comparative Study of Automatic Face Verification Algorithms on the BANCA Database", *Proc. Audio- and Video-Based Biometric Person Authentication*, Guildford, 2003, pp. 35-43.

[32] F. Samaria, *Face Recognition Using Hidden Markov Models*, PhD Thesis, University of Cambridge, 1994.

[33] C. Sanderson and K.K. Paliwal, "Fast Features for Face Authentication Under Illumination Direction Changes", *Pattern Recognition Letters*, Vol. 24, No. 14, 2003, pp. 2409-2419.

[34] C. Sanderson and K.K. Paliwal, "Identity Verification Using Speech and Face Information", *Digital Signal Processing*, in press.

[35] C. Sanderson, S. Bengio, “Augmenting Frontal Face Models for Non-Frontal Verification”, *Proc. Workshop Multimodal User Authentication*, Santa Barbara, 2003, pp. 165-172.

[36] C. Sanderson and S. Bengio, “Statistical Transformations of Frontal Models for Non-Frontal Face Verification”, *Proc. IEEE Int. Conf. Image Processing*, Singapore, 2004.

[37] M. Turk, A. Pentland, “Eigenfaces for Recognition”, *J. Cognitive Neuroscience*, Vol. 3, No. 1, 1991, pp. 71-86.

[38] J. D. Woodward, “Biometrics: Privacy’s Foe or Privacy’s Friend?”, *Proceedings of the IEEE*, Vol. 85, No. 9, 1997, pp. 1480-1492.

[39] J.L. Wayman, “Digital Signal Processing in Biometric Identification: a Review”, *Proc. IEEE Int. Conf. Image Processing*, Rochester, 2002, Vol. 1, pp. 37-40.

[40] J. Zhang, Y. Yan, M. Lades, “Face recognition: Eigenfaces, elastic matching, and neural nets”, *Proc. IEEE*, Vol. 85, No. 9, 1997, pp. 1422-1435.