User Authentication via Adapted Statistical Models of Face Images

Fabien Cardinaux [a], Conrad Sanderson [b], Samy Bengio [a]
[a] IDIAP Research Institute, Rue du Simplon 4, CH-1920 Martigny, Switzerland
[b] National ICT Australia (NICTA), Locked Bag 8001, Canberra 2601, Australia
{*cardinau, bengio*}*@idiap.ch, conradsand@ieee.org*

***Abstract*— It has been previously demonstrated that systems based on local features and relatively complex statistical models, namely 1D Hidden Markov Models (HMMs) and pseudo-2D HMMs, are suitable for face recognition. Recently, a simpler statistical model, namely the Gaussian Mixture Model (GMM), was also shown to perform well. In much of the literature devoted to these models, the experiments were performed with controlled images (manual face localization, controlled lighting, background, pose, etc). However, a practical recognition system has to be robust to more challenging conditions. In this article we evaluate, on the relatively difficult BANCA database, the performance, robustness and complexity of GMM and HMM based approaches, using both manual and automatic face localization. We extend the GMM approach through the use of local features with embedded positional information, increasing performance without sacrificing its low complexity. Furthermore, 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. Considerably more precise models can be obtained through the use of Maximum *a Posteriori* (MAP) training. We also show that face recognition techniques which obtain good performance on manually located faces do not necessarily obtain good performance on automatically located faces, indicating that recognition techniques must be designed from the ground up to handle imperfect localization. Finally, we show that while the pseudo-2D HMM approach has the best overall performance, authentication time on current hardware makes it impractical. The best trade-off in terms of authentication time, robustness and discrimination performance is achieved by the extended GMM approach.**

***Index Terms*— biometrics, access control, face recognition, face localization, Hidden Markov Models, Gaussian Mixture Models, local features, Maximum *a Posteriori* (MAP) training.**

I. Introduction

Biometric person recognition involves the use of inherent physiological characteristics of humans, such as faces, speech, iris patterns and fingerprints. Applications include surveillance, forensics, transaction authentication, and various forms of access control, such as immigration checkpoints and access to digital information [1], [23], [29], [44].

There are three distinct configurations of how a biometric recognition system can be used: the closed set identification task, the open set identification task, and the authentication task (also known as verification). In the closed set identification task, the job is to forcefully classify a given biometric sample as belonging to one of K persons (here K is the number of *known* persons). In the open set identification task, the task is to assign the given sample into one of K+1 classes (where the extra class represents an "unknown" or "previously unseen" person). Finally, in the authentication task the classifier assigns a given sample into one of two classes: either the sample belongs to a specific person, or it doesn't. In an access control scenario this translates to a person claiming an identity and providing a biometric sample to support this claim; the authentication system then classifies the person as either a true claimant or as an impostor.

The authentication task represents operation in an unconstrained environment, where *any* person/pattern could be encountered [19]. This is in contrast to the closed set identification task, where it is assumed that all the persons that are going to be encountered are already known. Further introductory and review material about the biometrics field can be found in the following papers: [12], [29], [39], [43], [44].

In this paper we exclusively focus on authentication based on face images. 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 [29]. Many techniques have been proposed for face classification; some examples are systems based on Principal Component Analysis (PCA) feature extraction [41], modular PCA [30], Elastic Graph Matching (EGM) [10], [21], and Support Vector Machines [34], Examples specific to statistical models include 1D Hidden Markov Models (HMMs) [35], pseudo-2D HMMs [13], [27] and Gaussian Mixture Models (GMMs) [5], [24], [38] (which can be considered as a simplified version of HMMs[1]). As an in-depth review of face recognition literature is beyond the scope of this paper, the interested reader is directed to the following review articles: [6], [18], [20], [46].

Statistical 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) [17] or DCTmod2 [38] is usually applied to each block, resulting in a set of feature vectors. In an analogous manner, 2D Gabor wavelets [22] can also be used.

In HMM based approaches, the spatial relation between

[1] Specifically, a GMM can be considered to be a single-state HMM, or a type of a multi-state ergodic HMM [32], where each state is modeled by a single gaussian.

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 [5] and to out-of-plane rotations [36]. 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 [13], [27], [35], [38], statistical models are trained using the Maximum Likelihood (ML) criterion via the Expectation Maximization (EM) algorithm [8]. It is generally known that one of the drawbacks of training via this paradigm is that large datasets are 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.* [13] proposed to use a well trained generic (non-person specific) model as the starting point for ML training. While the results in [13] were promising, they were obtained on the rather easy Olivetti Research Ltd. (ORL) database [35]. 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 [16], which can effectively circumvent the small training dataset problem.

Furthermore, we show that the performance of the overall face authentication system can be *highly dependent* on the performance of the face locator (detection) algorithm (i.e. the algorithm's ability to accurately locate a face, with no clipping or scaling problems). In other words, face classification techniques which obtain good performance on manually located faces do not necessarily obtain good performance on automatically located faces. We make the claim that the face classification technique *must be designed* from the ground up to handle imperfectly located faces.

Finally we show that complexity of a face classification system is an important consideration in a practical implementation. By "complexity" we mean the number of parameters to store for each person as well as the time required to make an authentication. If a face model is to be stored on an electronic card (e.g. an access card), the size of the model becomes an important issue. Moreover, the time needed to authenticate a person should not be cumbersome, implying the need to use techniques which are computationally simple.

The rest of this paper is organized as follows. Classifiers based on GMMs, 1D HMMs and P2D HMMs are described in Section II. An overview of the automatic face locator used in the experiments is given in Section III. Section IV covers preprocessing and feature extraction, while Section V provides a description of the BANCA database and its experiment protocols. Section VI is devoted to experiments involving the different training strategies, manual and automatic face localization, as well as effects of reducing the number of training images; the complexity of the models is also given. Conclusions and future areas of research are given in Section VII.

II. Classifiers Based on Statistical Models

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 T feature vectors $X = \{\mathbf{x}_t\}_{t=1}^{T}$ supporting the claim (extracted from the given face), we find an opinion on the claim using:

$$\Lambda(X) = \log P(X|\lambda_C) - \log P(X|\lambda_{\overline{C}}) \quad (1)$$

where $P(X|\lambda_C)$ is the likelihood of the claim coming from the true claimant and $P(X|\lambda_{\overline{C}})$ is an approximation of 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* [25], [33]; it is typically trained using data from many people. 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 [8], [11].
2) ML training with a generic (non-client specific) model as the starting point (as in [13]); data from many people are 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. (1) for all generative approaches.
3) MAP training [16]; 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, $\lambda_{\mathtt{MAP}}$, is defined as:

$$\begin{aligned} \lambda_{\mathtt{MAP}} &= \arg\max_{\lambda} P(\lambda|X) && (2) \\ &= \arg\max_{\lambda} P(X|\lambda)P(\lambda) && (3) \end{aligned}$$

Assuming λ to be uniform is equivalent to having a non-informative $P(\lambda)$, reducing the solution of $\lambda_{\mathtt{MAP}}$ 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 II-A.

A. Gaussian Mixture Model

In the GMM approach, all feature vectors are assumed to be independent. The likelihood of a set of feature vectors is found with

$$P(X|\lambda) = \prod_{t=1}^{T} P(\mathbf{x}_t|\lambda) \quad (4)$$

where

$$\begin{aligned} P(\mathbf{x}|\lambda) &= \sum_{k=1}^{N_G} m_k \, \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) && (5) \\ \lambda &= \{m_k, \mu_k, \mathbf{\Sigma}_k\}_{k=1}^{N_G} && (6) \end{aligned}$$

Here, $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$ is a D-dimensional gaussian density function [11] with mean μ and diagonal covariance matrix $\mathbf{\Sigma}$. N_G is the number of gaussians and m_k is the weight for gaussian k (with constraints $\sum_{k=1}^{N_G} m_k = 1$ and $\forall\ k : m_k \geq 0$).

An implementation of MAP training for client model adaptation consists of using a global parameter to tune the relative importance of the prior. In this case, the equations for adaptation of the parameters are [16], [25], [33]:

$$\hat{w}_k = \left[\alpha w_k + (1-\alpha) \sum_{t=1}^{T} P(k|\mathbf{x}_t) \right] \gamma \tag{7}$$

$$\hat{\mu}_k = \alpha \mu_k + (1-\alpha) \frac{\sum_{t=1}^{T} P(k|\mathbf{x}_t)\mathbf{x}_t}{\sum_{t=1}^{T} P(k|\mathbf{x}_t)} \tag{8}$$

$$\hat{\mathbf{\Sigma}}_k = \alpha \left(\mathbf{\Sigma}_k + \mu_k \mu_k^{\prime} \right) + (1-\alpha) \frac{\sum_{t=1}^{T} P(k|\mathbf{x}_t)\mathbf{x}_t\mathbf{x}_t^{\prime}}{\sum_{t=1}^{T} P(k|\mathbf{x}_t)} - \hat{\mu}_k \hat{\mu}_k^{\prime} \tag{9}$$

where $\hat{w}_k$, $\hat{\mu}_k$ and $\hat{\mathbf{\Sigma}}_k$ are respectively the new weight, mean and covariance matrix of the k-th gaussian, w_k, μ_k and $\mathbf{\Sigma}_k$ are the corresponding parameters in the generic model, $P(k|\mathbf{x}_t)$ is the posterior probability of the k-th gaussian (from the client model from the previous iteration), $\alpha \in [0, 1]$ is the adaptation factor chosen empirically on a separate validation set, and finally γ is computed over all adapted weights to ensure they sum to unity. Each $\hat{\mathbf{\Sigma}}_k$ is forced to be diagonal by setting the off-diagonal elements to zero. Note that in Eqn. (7) the new mean is simply a weighted sum of the prior mean and new statistics; $(1-\alpha)$ can hence be interpreted as the amount of faith we have in the new statistics.

The above formulation of MAP training makes the assumption of independence between the parameters of the individual mixture components and the set of the mixture weights; furthermore we consider that we can model the prior knowledge about the parameter vector of mixture weights with a Dirichlet density and the prior knowledge about the means and variances with normal-Wishart densities [16].

The adaptation procedure is iterative, thus an initial client model is required; this is accomplished by copying the generic model. It has been observed that it is sometimes preferable to adapt only the means of the gaussians [33]; we will empirically show that this is also valid for our experiments in Section VI. When only the means are adapted the other parameters are copied from the generic model.

1) Embedding Positional Information: If each feature vector in the set X describes a different part of the face, then a classifier based on GMMs effectively loses the spatial relations between face parts. As the spatial relations can carry discriminatory information, we propose to increase the performance of the GMM approach (without sacrificing its simplicity) by restoring a degree of spatial relations via embedding positional information into each feature vector. Doing so should place a weak constraint on the areas that each gaussian in the GMM can model, thus making a face model more specific. Formally, an extended feature vector for position (a, b) is obtained with:

$$\mathbf{x}_{(a,b)}^{\texttt{extended}} = \begin{bmatrix} \mathbf{x}_{(a,b)}^{\texttt{original}} \\ a \\ b \end{bmatrix}$$

where $\mathbf{x}_{(a,b)}^{\texttt{original}}$ is the original feature vector for position (a, b). We shall refer to a GMM system using extended feature vectors as GMMext.

B. 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 [35] or left-right HMM in the context of speech recognition [32]. Here the face is represented as a sequence of overlapping *rectangular* blocks from top to bottom of the face (see Fig. 1 for an example). 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 $B = \{b_j(\mathbf{x}_t)\}$, where
$$b_j(\mathbf{x}_t) = p(\mathbf{x}_t | q_t = S_j) \tag{10}$$
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:

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

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:

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

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

The calculation of this likelihood according to the direct definition in Eqn. (13) involves an exponential number of computations. In practice the Forward-Backward procedure is used [32]; it is mathematically equivalent, but considerably more efficient.

For the case of the 1D HMM, MAP adaptation of the means is [c.f. Eqn. (8)]:

$$\hat{\mu}_{k,i} = \alpha \mu_{k,i}^{w} + (1-\alpha) \frac{\sum_{t=1}^{T} P(q_t = i|\mathbf{x}_t) P(m_t^i = k|\mathbf{x}_t)\mathbf{x}_t}{\sum_{t=1}^{T} P(q_t = i|\mathbf{x}_t) P(m_t^i = k|\mathbf{x}_t)} \tag{14}$$

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

Compared to the GMM approach described in Section II-A, the spatial constraints are much more strict, mainly due to the rigid preservation of horizontal spatial relations (e.g. horizontal positions of the eyes). The vertical constraints are relaxed, though they still enforce the top-to-bottom segmentation (e.g. the eyes have to be above the mouth). The non-rigid constraints allow for a degree of vertical translation and some vertical stretching (caused, for example, by an imperfect face localization).

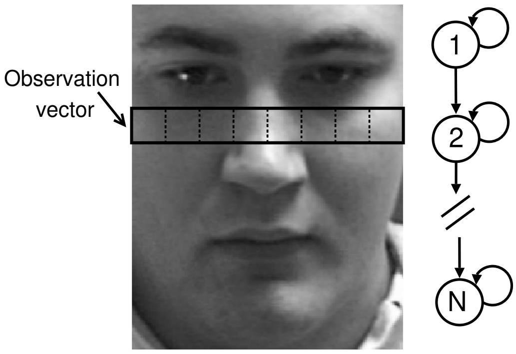

Fig. 1. Sampling window and 1D HMM topology.

C. 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 the "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 [13], [35] and the training process is described in detail in [28]. As shown in Fig. 2, 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 the bottom (forehead, eyes, nose, mouth). Note that the opposite choice has been made in [13], [35]. 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 corresponding equation for MAP adaptation of the means [c.f. Eqns. (8) and (14)] is:

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

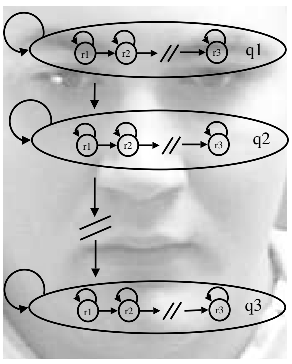

Fig. 2. P2D HMM: the emission distributions of the vertical HMM are estimated by horizontal HMMs. q_i represent the states of the main HMM and r_j represent the embedded HMMs states.

with:

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

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

The degree of spatial constraints present in the P2D HMM approach can be thought of as being somewhere in between the GMM and the 1D HMM approaches. While the GMM approach has no spatial constraints and the 1D HMM has rigid horizontal constraints, the P2D HMM approach has relaxed constraints in both directions. However, the constraints still enforce the left-to-right segmentation of the embedded HMMs (e.g. the left eye has to be before the right eye), and top-to-bottom segmentation (e.g. like in the 1D HMM approach, the eyes have to be above the mouth). The non-rigid constraints allow for a degree of both vertical and horizontal translations, as well as some vertical and horizontal stretching of the face.

III. Face Localization

Face recognition results in the literature are usually presented assuming manual face localization (e.g. see [13], [27], [28], [35]); in only relatively few publications performance evaluation is found while using automatic face localization (e.g. [5], [34]). While assuming manual (i.e. perfect) localization makes the results independent of the quality of the face localization system, they are optimistically biased compared to a real life system, where the face needs to be automatically located. There is no guarantee that the automatic face localization system will provide a correctly located face (i.e. the face may be translated and/or at an incorrect scale). In this paper we present results for both manually and imperfectly located faces.

For "manual face localization" experiments, we use the manually annotated eye center positions. For "automatic face localization" experiments, we use the face detector[2] proposed by Fröba and Ermst in [15]. The detector employs local features based on the *Modified Census Transform*, which represent each location of the image by a binary pattern computed from a $3{\times}3$ pixel neighbourhood. Face detection is carried out by analyzing all possible windows in the given image at different scales; each window is classified as either containing a face or the background. The classification is performed by a cascade classifier similar to the approach proposed by Viola and Jones [42]; training of the classifier is accomplished using a version of the boosting algorithm [14]. In our experiments the eye positions are inferred from the position and scale of the window with the best score at the last stage of the classifier. Note that this assumes that at most only one face is present in each image.

If all the windows were classified as containing the background, we consider that the given image does not contain a face and we perform the authentication using, if available, other images supporting the claim. If all given images are

[2] A recent survey of face localization/detection methods is given in [45].

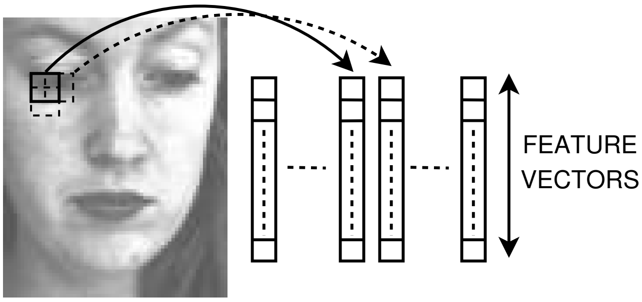

Fig. 3. Conceptual example of block by block image analysis.

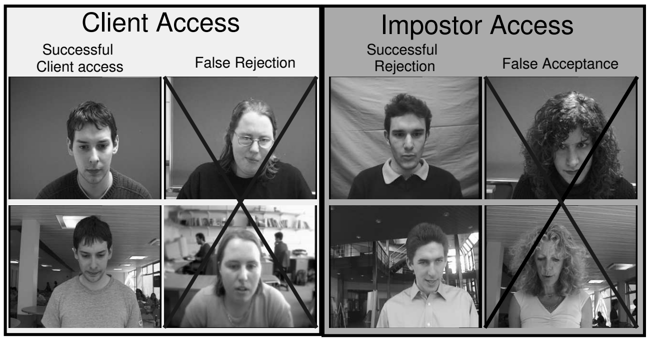

Fig. 4. Example correct and incorrect verification from the BANCA database. Top row contains training images (from the controlled condition) while the bottom row contains test images from degraded and adverse conditions.

deemed not to contain a face, the claim is considered to have come from an impostor.

IV. Pre-processing and Feature Extraction

Based on given eye positions, a gray-scale 80×64 (rows $\times$ columns) face window is cropped out of each valid image (i.e. an image which is deemed to contain a face). When using manually found eye positions, 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). Fig. 1 shows an example face window.

Histogram equalization is used to normalize the face images photometrically. We then extract DCTmod2 features from each image face [38]. We have found this combination of histogram equalization and feature extraction to provide good results in preliminary experiments. The feature extraction process is summarized as follows. The face window is analyzed on a block by block basis; each block is $N_P \times N_P$ (here we use N_P=8) and overlaps neighbouring blocks by a configurable amount of pixels[3]. Fig. 3 illustrates such a block by block decomposition.

Each block is decomposed in terms of 2D Discrete Cosine Transform (DCT) basis functions [17]. A feature vector for a block located at row a and column b is then constructed as:

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

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; the deltas are computed using DCT coefficients extracted from neighbouring blocks. Compared to traditional DCT feature extraction [13], [27], the first three DCT coefficients are replaced by their respective deltas in order to reduce the effects of illumination direction changes, without losing discriminative information. In this study we use M=15 (based on [38]), resulting in an 18 dimensional feature vector for each block. The degree of overlap has three main effects:

1) As the delta coefficients are computed from neighbouring blocks, the larger the overlap between the blocks, the smaller the spatial area used to derive each feature vector.
2) With a large overlap, the DCT coefficients from a set of (horizontally or vertically) consecutive blocks will not vary abruptly.
3) When using a large overlap, the parts of each face are in effect "sampled" at various degrees of translations, resulting in models which should be robust to minor translations of the faces. This is in *addition* to the translation robustness provided by the GMM classifier, where the location of each block has little influence. By itself, GMM's built-in robustness only works when the size of the translation is equivalent to an integral multiple of the block size.

[3]A similar overlapping approach is used in the processing of speech signals [31], [33], [40].

V. BANCA Database and Experiment Protocols

The multi-lingual BANCA database [2] was designed to evaluate multi-modal identity authentication with various acquisition devices under several scenarios. The database is comprised of four separate corpora, each containing 52 subjects; the corpora are named after their country of origin. Each subject participated in 12 recording sessions in different conditions and with different cameras. Each of these sessions contains two video recordings: one true claimant access and one impostor attack. Five "frontal" (not necessarily directly frontal) face images have been extracted from each video recording. Sessions 1-4 contain images for the *controlled* condition, while sessions 5-8 and 9-12 respectively contain *degraded* and *adverse* conditions. The latter two conditions differ from the *controlled* condition in terms of image quality, lighting, background and pose. See Fig. 4 for an example of the differences.

According to the original experiment protocols, there are seven distinct configurations that specify which images can be used for training and testing: Matched Controlled (Mc), Matched Degraded (Md), Matched Adverse (Ma), Unmatched Degraded (Ud), Unmatched Adverse (Ua), Pooled test (P) and Grand test (G). Table I 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

TABLE I

USAGE OF THE SEVEN BANCA PROTOCOLS (C: CLIENT, I: IMPOSTOR). THE NUMBERS REFER TO THE ID OF EACH SESSION.

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

four different scenarios should also make the results easier to interpret.

According to the BANCA experiment protocols, experiments should be performed on each corpus independently. The protocols further dictate that the subjects in each corpus are equally split into the validation and test sets. Subjects in the validation set are used to optimize the authentication system (e.g. to find the optimal number of gaussians and the decision threshold), while subjects from the test set are used for final performance evaluation. Note that this amounts to using only 26 subjects in the final stage. To increase the number of subjects, we merged the English and French corpora, resulting in a total of 104 subjects. In a similar manner to the original protocols, the resulting population was then divided into two groups of 52 subjects.

Authentication systems make two types of errors: a False Acceptance (FA), which occurs when the system accepts an impostor, or a False Rejection (FR), which occurs when the system refuses a true claimant. The performance is generally measured in terms of False Acceptance Rate (FAR) and False Rejection Rate (FRR), defined as:

$$\text{FAR} = \frac{\text{number of FAs}}{\text{number of impostor accesses}} \quad (17)$$

$$\text{FRR} = \frac{\text{number of FRs}}{\text{number of true claimant accesses}} \quad (18)$$

The FAR and FRR are usually related, meaning that decreasing one usually increases the other.

To aid the interpretation of performance, the two error measures are often combined using the Half Total Error Rate (HTER[4]), defined as:

$$\text{HTER}(\tau, \mathcal{D}) = \frac{\text{FAR}(\tau, \mathcal{D}) + \text{FRR}(\tau, \mathcal{D})}{2} \quad (19)$$

where $\text{FAR}(\tau, \mathcal{D})$ and $\text{FRR}(\tau, \mathcal{D})$ are the FAR and FRR, respectively, for a decision threshold τ and dataset $\mathcal{D}$. A particular case of the HTER, known as the Equal Error Rate (EER), occurs when the threshold is adjusted so that FAR=FRR on a particular dataset $\mathcal{D}'$ (which could be different from $\mathcal{D}$).

In some situations it may be more important to have a system with a very small FAR, while in other situations a small FRR might be more useful. In order to see performance with respect to the trade-off between the FAR and FRR, Receiver Operating Characteristics (ROC) and the Detection Error Tradeoff (DET) curves [26] are often used. However, it has been recently observed that these curves can be misleading [4] as they do not take into account that, in real life, the threshold has to be selected *a priori*. In this paper we use the *Expected Performance Curve* (EPC) [4], which is in coherence with the original BANCA protocols and can be interpreted as an unbiased version of the ROC curve.

Let $\omega \in [0, 1]$ reflect the trade-off between the FAR and the FRR. In the EPC approach, an optimal threshold $\tau_{\texttt{opt}}$ is found for various values of ω as follows:

$$\tau_{\texttt{opt}} = \arg\min_{\tau} \ \omega \text{FAR}\left(\tau, \mathcal{D}_{\texttt{valid}}\right) + (1 - \omega) \text{FRR}\left(\tau, \mathcal{D}_{\texttt{valid}}\right) \quad (20)$$

where $\mathcal{D}_{\texttt{valid}}$ is the validation dataset. The HTER (using $\tau_{\texttt{opt}}$) is then calculated on the test set ($\mathcal{D}_{\texttt{test}}$) and is plotted as a function of ω. For ω=0.5, the above procedure is equivalent to finding the minimum EER on the validation set, and then calculating the HTER on the test set using an *a priori* threshold.

VI. 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 image. 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 and French corpora), 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; experiments on the validation set showed that an overlap of four pixels is better for the GMM approaches while an overlap of seven pixels is preferred by the P2D HMM approach. For the 1D HMM approach, a seven pixel overlap was also used, but feature vectors from the same row of blocks were concatenated to form a large observation vector. To keep the dimensionality of the resultant vector reasonable, we chose to concatenate vectors 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 to select the size of the model (e.g. number of states and gaussians) as well as other hyper-parameters, such as the adaptation coefficient α, and the decision threshold τ. The hyper-parameters were chosen to minimize the EER. The final performance of each model was then found on the test set in terms of HTER (and/or EPCs, where applicable).

It has been observed that in applications such as speaker authentication [25], [33], MAP based training obtains best performance when only the means are adapted (rather than also adapting the covariance matrices and weights). Fig. 5 shows EPCs for the standard GMM based system for three cases: (i) all parameters are adapted, (ii) means and covariance matrices are adapted, (iii) only means are adapted. Database protocol P was employed in this evaluation. As adapting only the means provides the best performance, we have elected to use this strategy for both GMM and HMM approaches[5]. Hence

[4]The HTER can be thought of as a special case of the Decision Cost Function (DCF) [3], [9].

[5]Computational limitations and time constraints prevented us from repeating this experiment for the HMM based approaches.

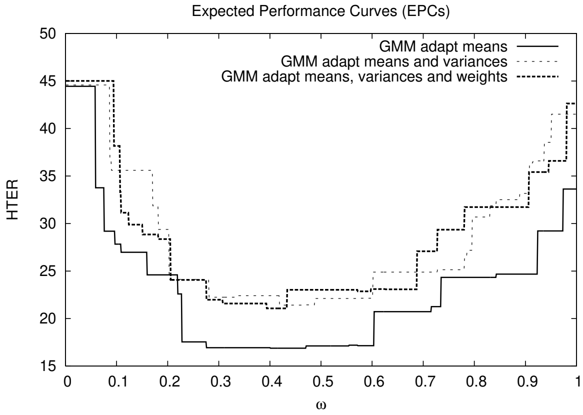

Fig. 5. EPC performance of standard GMM based system trained via MAP adaptation. Three configurations of MAP adaptation are shown.

TABLE II

OPTIMAL PARAMETERS FOR SYSTEMS BASED ON GMM (STANDARD FEATURES), GMMEXT (EXTENDED 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.

System	Number of states		Gaussians per state	Total gaussians
	main HMM	embedded HMM		
GMM *ML*	-	-	-	256
GMM *init*	-	-	-	512
GMM *adapt*	-	-	-	512
GMMext *ML*	-	-	-	256
GMMext *init*	-	-	-	512
GMMext *adapt*	-	-	-	1024
1D HMM *ML*	16	-	1	16
1D HMM *init*	32	-	1	32
1D HMM *adapt*	32	-	1	32
P2D HMM *ML*	8	16	4	512
P2D HMM *init*	16	16	2	512
P2D HMM *adapt*	16	4	64	4096

for the rest of this paper, the MAP training strategy will refer to the adaptation of the means only.

Throughout the remainder of this paper, the following notation is used. GMM indicates the GMM approach with standard DCTmod2 feature vectors, while GMMext indicates the GMM approach with extended DCTmod2 feature vectors; 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*.

Table II shows the optimal number of states and gaussians per state for the HMM approaches, as well as the total number of gaussians for all approaches. It can be observed that MAP training generally allows the total number of gaussians to be higher (thus modeling the faces more accurately), when compared to the two ML based training paradigms. The P2D HMM approach utilizes the largest number of gaussians, followed by the GMMext approach.

For comparison purposes, we also evaluate the performance of a PCA based system, which in effect has rigid constraints between face parts. 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 small size of the client specific training dataset, and since PCA feature extraction results in one feature vector per face, 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 [36], [39]. Feature vectors with 160 dimensions were found to provide optimal performance on the validation set.

In Section VI-A we present the results for manual face localization, while Section VI-B contains results for imperfect and automatic face localization. In Section VI-C we study the effects of varying the number of training images and finally in Section VI-D we compare the complexity of the local feature approaches.

Note that the result tables presented in Sections VI-A and VI-B also contain performance figures for the two best systems reported in [34]. The first system is based on combination of Linear Discriminant Analysis and Normalized Correlation (LDA/NC), while the second system is based on a Support Vector Machine (SVM) classifier. Like the PCA based system, these LDA/NC and SVM systems are holistic in nature. It must be noted that in [34], only the English corpus was used and a different automatic face localization system was employed. As such the results from [34] 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 manually located faces).

A. Manual Face Localization

Table IV(a) shows the results in terms of HTER for manual face localization; Fig. 6 shows the corresponding EPCs. When the different training strategies are compared, MAP training provides a clear performance advantage in almost all the cases. The only exception is the 1D HMM approach for which all training approaches obtain similar performance. 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).

When the performance across different models is compared, it can be seen that the two HMM approaches (1D and P2D HMM) obtain considerably better performance than the two GMM based approaches. Comparing the standard GMM and the GMMext approach, the results show that use of extended feature vectors can result in better performance; while this is most pronounced when using ML based training, the performance differences are small when using MAP training.

The 1D HMM outperforms the P2D HMM approach when ML training is utilized; this can be explained by the inherently much larger number of parameters used in P2D HMM (hence requiring a larger training dataset). However, when MAP training is used, the small dataset problem is effectively circumvented, resulting in the P2D HMM approach obtaining the best overall performance.

B. Imperfect and Automatic Localization

Prior to using the automatic face locator described in Section III, we first study how each system is affected by

TABLE III

HTER PERFORMANCE FOR (A) **manual face localization**, AND (B) **automatic face localization**, USING GMM (STANDARD FEATURES), GMMEXT (EXTENDED 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. "$*$" INDICATES THE BEST RESULT FOR A PROTOCOL, WHILE BOLDFACE INDICATES THE BEST RESULT WITHIN A MODEL TYPE AND PROTOCOL. SEE THE TEXT FOR A NOTE ON THE RESULTS FROM [34].

System	Protocol			
	Mc	Ud	Ua	P
PCA	9.5	20.9	20.8	18.4
LDA/NC (from [34])	4.9	16.0	20.2	14.8
SVM (from [34])	5.4	25.4	30.1	20.3
GMM *ML*	12.9	28.9	26.0	22.9
GMM *init*	12.8	29.7	28.3	23.8
GMM *adapt*	**8.9**	**17.3**	**20.9**	**17.0**
GMMext *ML*	11.2	24.5	24.4	20.8
GMMext *init*	10.5	24.3	24.7	20.8
GMMext *adapt*	**8.5**	**17.5**	**20.8**	**16.4**
1D HMM *ML*	9.1	17.8	17.1	15.9
1D HMM *init*	9.1	**15.6**	17.4	**14.7**
1D HMM *adapt*	**6.9**	16.3	**17.0**	**14.7**
P2D HMM *ML*	9.0	19.0	18.0	17.5
P2D HMM *init*	8.6	16.5	19.2	17.0
P2D HMM *adapt*	$*$ **4.6**	$*$ **15.3**	$*$ **13.1**	$*$ **13.5**

(a) HTER performance for **manual face localization**

System	Protocol			
	Mc	Ud	Ua	P
PCA	22.4	29.7	33.7	29.0
LDA/NC (from [34])	22.6	25.4	27.1	25.2
SVM (from [34])	19.7	30.4	33.2	27.8
GMM *ML*	16.7	33.3	33.3	27.7
GMM *init*	19.8	35.0	35.1	29.7
GMM *adapt*	**9.5**	**21.0**	**24.8**	**19.5**
GMMext *ML*	15.8	27.7	29.3	24.9
GMMext *init*	17.5	31.9	30.4	27.2
GMMext *adapt*	**8.5**	**18.4**	**22.5**	**19.1**
1D HMM *ML*	21.0	28.8	29.5	27.0
1D HMM *init*	21.3	30.1	31.4	28.1
1D HMM *adapt*	**13.8**	**25.9**	**23.4**	**21.7**
P2D HMM *ML*	12.1	25.2	26.9	22.3
P2D HMM *init*	13.5	24.6	26.5	22.5
P2D HMM *adapt*	$*$ **6.5**	$*$ **15.9**	$*$ **14.7**	$*$ **14.7**

(b) HTER performance for **automatic face localization**

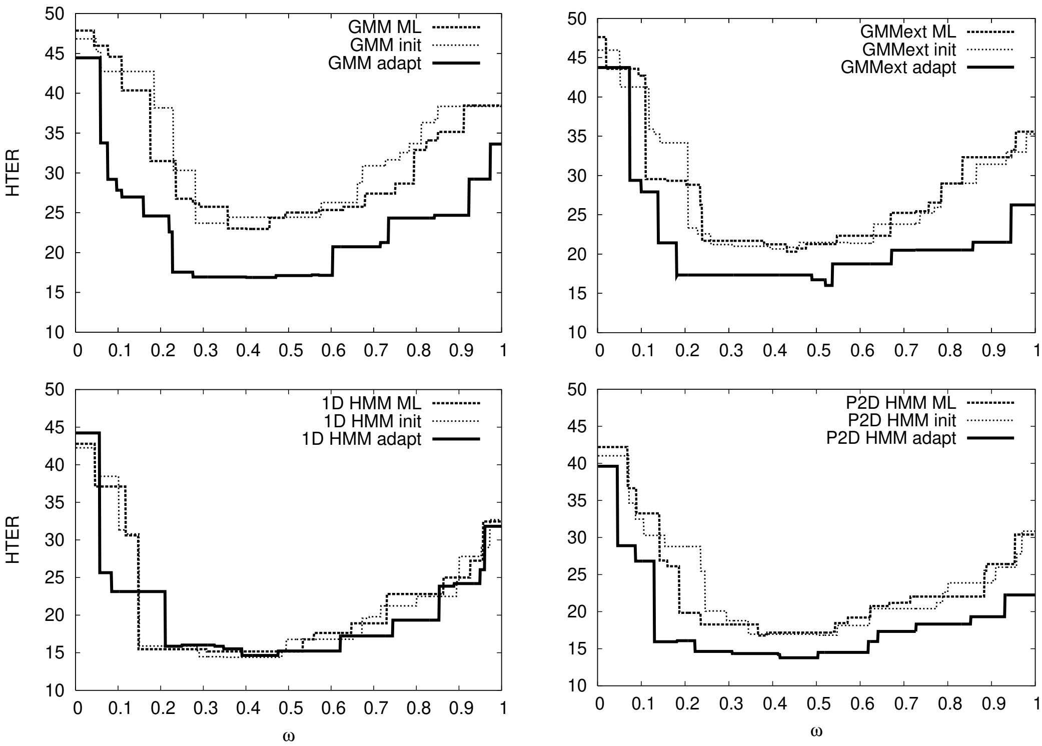

Fig. 6. EPCs for manual face localization, for the GMM, GMMext 1D HMM and P2D HMM systems, using three different methods: ML, init, and adapt (see caption of Table III for more details).

an increasing amount of error in the position of the eyes. For this set of experiments we used exactly the same models as in Section VI-A (i.e. trained with manually localized faces). The eye positions were artificially perturbed using:

$$eye_x = eye_x^{\mathtt{gt}} + \xi_x \quad (21)$$

$$eye_y = eye_y^{\mathtt{gt}} + \xi_y \quad (22)$$

where $eye_x^{\mathtt{gt}}$ and $eye_y^{\mathtt{gt}}$ are the ground-truth (original) coordinates for an eye. ξ_x and ξ_y are random variables which follow a normal distribution such that $\xi \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = V \cdot D_{\mathtt{eyes}}$, with $D_{\mathtt{eyes}}$ being the Euclidean distance between the two eyes. $V \in [0, 1]$ and can be interpreted as the amount of introduced error in the face location.

Results in Fig. 7 show that GMM, GMMext and P2D HMM based systems are quite robust to imperfect face localization.

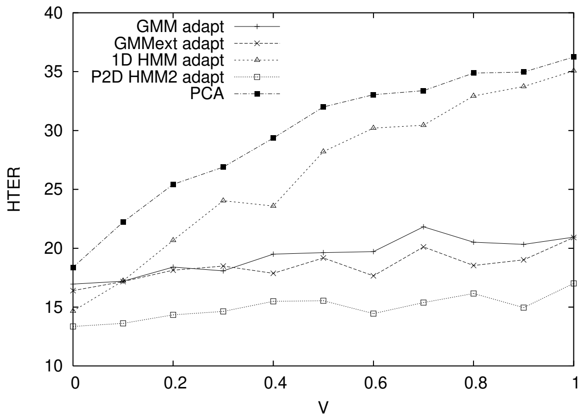

Fig. 7. Performance for an increasing amount of error in eye locations.

In contrast, the PCA and 1D HMM systems are significantly more sensitive, with their discrimination performance rapidly decreasing as V is increased. We attribute this performance degradation to the more constrained spatial relation between face parts; while the 1D HMM system allows for some vertical displacement, it has rigid constraints in the horizontal direction; in the PCA based system the relations are rigidly preserved along both axes.

Table IV(b) shows that the observations from perturbation experiments are confirmed when the automatic face locator is utilized. The PCA system is the most affected, followed by the 1D HMM. In Table IV(a) it was shown that when using MAP based training and manual face localization, the 1D HMM approach outperforms the two GMM based systems; however, for automatic face localization, the GMMext approach outperforms the 1D HMM system. We also note that the spatial constraints present in the GMMext approach do not affect the robustness of the system. The P2D HMM system again obtains the best overall performance, with minimal degradation in discrimination ability when compared to manually located faces.

C. Number of Training Images

The relatively small number of face images available to train each client model can be a limiting factor in obtaining precise face models. In some applications, such as surveillance, there may be only one reference image (e.g. a passport photograph). In the experiments reported in Sections VI-A and VI-B, five images were available for each client; the number of images was artificially increased to ten by mirroring each original image. In this section we evaluate the effects of decreasing the number of original training images.

Fig. 8 shows the performance as a function of the number of original images (i.e. mirrored versions were also utilized). Database protocol P was employed in this evaluation. Irrespective of the training strategy and model, the greatest improvement generally occurs when two training images are utilized instead of one; moreover, discrimination performance tends to saturate at three images. The exception is the MAP trained P2D HMM approach, where there is no clear benefit in utilizing more than one image. Overall, MAP training is the least sensitive to the number of training images. Lastly, the GMM and GMMext systems benefit the most from an increase in the number of training images.

D. 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. an access 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 IV shows the complexity of each local feature model used in our experiments (using hyper-parameters tuned for optimal discrimination performance, 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 speed, we believe the presented timings are indicative.

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. For example, when using the GMM approach and an equal number of gaussians for both ML and MAP training, the number of client specific parameters for MAP trained models is about half of the number required for ML based training.

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. The GMM, GMMext and 1D HMM approaches have short training and authentication times of around three and one seconds, respectively. We note that for these three approaches, the pre-processing steps considerably penalize the speed of the authentication.

When using MAP trained models, the P2D HMM approach has a considerably higher training and authentication time,

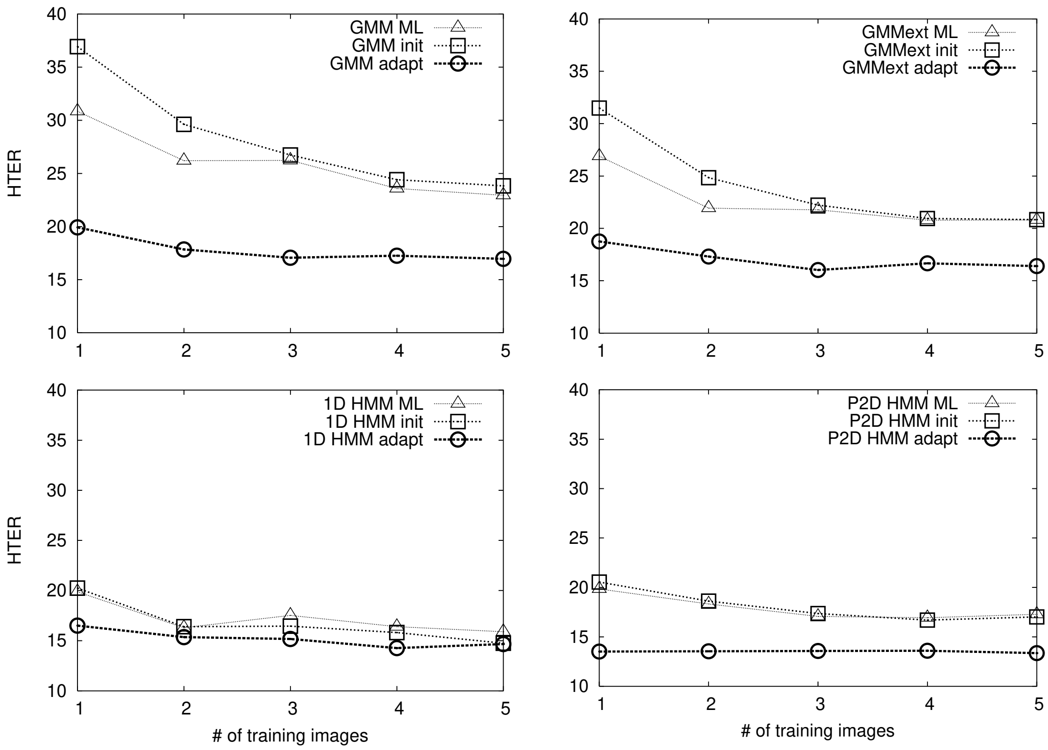

Fig. 8. Performance as a function of the number of original training images.

TABLE IV

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).

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	9,472	18,944	9,216	10,496	20,992	20,480	4,063	8,127	4,032	19,207	19,471	73,728
world model training time	295s (163s)	470s (337s)	470s (355s)	253s (120s)	364s (231s)	679s (546s)	181s (3s)	184s (6s)	192s (14s)	2873s (2695s)	1873s (1695s)	7967s (7789s)
client model training time	2s (0.5s)	2s (1s)	2s (1s)	2s (0.5s)	2s (1s)	3s (1.5s)	2s (0.5s)	2s (0.5s)	3s (2s)	65s (64s)	88s (87s)	251s (250s)
time for authentication of one claim (5 images)	0.95s (0.07s)	1.10s (0.22s)	1.12s (0.24s)	0.93s (0.05s)	1.02s (0.14s)	1.28s (0.40s)	1.22s (0.13s)	1.25s (0.16s)	1.31s (0.22s)	5.74s (4.65s)	7.25s (6.16s)	19.89s (18.80s)

at approximately 4 minutes for training each client model and 20 seconds for an authentication. With current computing resources, this authentication time can be considered as being too long for practical deployment purposes. When using ML trained models, the training and authentication time is significantly reduced, which is partly due to the total number of gaussians being smaller. However, ML trained models obtain considerably worse discrimination performance. Table IV(b) shows that the MAP trained GMMext approach outperforms the ML trained P2D HMM approach, suggesting that in practical terms the GMMext approach obtains the best trade-off in terms of authentication time, robustness and discrimination performance.

VII. Conclusions and Future Work

The findings of this paper can be summarized as follows:

- The traditionally used Maximum Likelihood (ML) training approach has problems estimating robust model parameters when there is only a few training images available. Using Maximum *a Posteriori* (MAP) based training results in considerably more precise models, leading to higher discrimination performance.
- Good performance on manually located faces does not necessarily reflect good performance in real life conditions, where an automatic localization system must be used. As automatic localization cannot guarantee perfect face localization, any new face classification technique must be designed from the ground up to handle imperfectly located faces.
- Systems that utilize rigid spatial constraints between face parts (such as PCA and 1D HMM based), are easily affected by face localization errors, which are caused by an automatic face locator. In contrast, systems which have relaxed constraints (such as GMM and P2D HMM based), are quite robust.
- While the 1D HMM based approach achieves promising performance for manually (i.e. perfectly) located faces and outperforms the extended GMM approach, for auto-

matically located faces its performance degrades considerably and is worse than the extended GMM approach.

- Use of feature vectors with embedded positional information somewhat increases the performance of the GMM approach, with no loss of robustness to errors in face localization. Along with the good performance of the P2D HMM approach, this indicates that spatial relations between face parts carry discriminative information.
- The P2D HMM approach is overall the most robust and obtains the best discrimination performance, when compared to the 1D HMM and GMM based approaches. However, it also the most computationally intensive approach, making it impractical for application use on current hardware.
- The best trade-off in terms of complexity, robustness and discrimination performance is achieved by the extended GMM approach.

Future research includes the following avenues:

- Currently in the extended GMM approach the degree of influence of positional information is not controlled; higher performance might be attained if more weight is placed on this information. A possible indirect approach to accomplish this is by placing an upper limit (during training) on the variances for the dimensions representing positional information.
- We conjecture that a major source of authentication errors is the pose mismatch between the training and test faces. An initial investigation on transforming frontal face models to represent non-frontal views is given in [37]. The results are encouraging, indicating there is room for improvement by reducing the pose mismatch.
- The MAP trained P2D HMM system could be deliberately detuned (e.g. by reducing the number of gaussians in each state) in order to reduce its complexity, and hence reduce the time taken to perform an authentication. This will probably come at the cost of a loss in discrimination performance, though the extent of this loss remains to be seen. Embedding positional information into the feature vectors may mitigate the loss.

Acknowledgments

The authors thank J. Mariéthoz for fruitful discussions and 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 [7]. National ICT Australia is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

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 (AVBPA)*, 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] S. Bengio, J. Mariéthoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication", *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, 2004, Toledo, pp. 279-284.

[5] 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 (AVBPA)*, Guildford, 2003, pp. 911-920.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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 (ICASSP)*, Orlando, 2002, Vol. IV, pp. 4060-4062.

[13] 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.

[14] Y. Freund, R.E. Shapire "A short introduction to boosting", *Journal of Japanese Society for Artificial Intelligence*, No. 14, 1999, pp. 771-780.

[15] B. Fröba, A. Ernst "Face Detection with the Modified Census Transform", *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (AFGR)*, Seoul, 2004, pp. 91-96.

[16] J.-L. Gauvain, 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.

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

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

[19] Behrooz Kamgar-Parsi, Behzad Kamgar-Parsi, A. Jain, J. Dayhoff, "Aircraft Detection: A Case Study in Using Human Similarity Measure", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 12, 2001, pp. 1404-1414.

[20] 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*, Vol 97, No. 1, 2005, pp. 103-135.

[21] 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.

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

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

[24] S. Lucey, T. Chen, "A GMM parts based face representation for improved verification through relevance adaptation", *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, Washington D.C., 2004, pp. 855-861.

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

[26] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, "The DET curve in assessment of detection task performance", *Proc. Eurospeech'97*, Rhodes, Greece, 1997, pp. 1895-1898.

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

[28] A. Nefian, M. Hayes, "Maximum likelihood training of the embedded HMM for face detection and recognition", *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Vancouver, 2000, Vol. 1, pp. 33-36.
[29] 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.
[30] 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.
[31] J. Picone, "Signal Modeling Techniques in Speech Recognition", *Proceedings of the IEEE*, Vol. 81, No. 9, 1993, pp. 1215-1247.
[32] 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.
[33] 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.
[34] 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 (AVBPA)*, Guildford, 2003, pp. 35-43.
[35] F. Samaria, *Face Recognition Using Hidden Markov Models*, PhD Thesis, University of Cambridge, 1994.
[36] C. Sanderson, S. Bengio, "Extrapolating Single View Face Models for Multi-View Recognition", *Proc. Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, Melbourne, 2004, pp. 581-586.
[37] C. Sanderson, S. Bengio, "Statistical Transformations of Frontal Models for Non-Frontal Face Verification", *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Singapore, 2004, pp. 585-588.
[38] C. Sanderson, K.K. Paliwal, "Fast Features for Face Authentication Under Illumination Direction Changes", *Pattern Recognition Letters*, Vol. 24, No. 14, 2003, pp. 2409-2419.
[39] C. Sanderson, K.K. Paliwal, "Identity Verification Using Speech and Face Information", *Digital Signal Processing*, Vol. 14, No. 5, 2004, pp. 449-480.
[40] F.K. Soong, A.E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 36, No. 6, 1988, pp. 871-879.
[41] M. Turk, A. Pentland, "Eigenfaces for Recognition", *J. Cognitive Neuroscience*, Vol. 3, No. 1, 1991, pp. 71-86.
[42] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2001, Vol. 1, pp. 511-518.
[43] J.L. Wayman, "Digital Signal Processing in Biometric Identification: a Review", *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Rochester, 2002, Vol. 1, pp. 37-40.
[44] J.D. Woodward, "Biometrics: Privacy's Foe or Privacy's Friend?", *Proceedings of the IEEE*, Vol. 85, No. 9, 1997, pp. 1480-1492.
[45] M.-H. Yang, D. Kriegman, N. Ahuja, "Detecting Faces in Images: A Survey", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, 2002, pp. 34-58.
[46] J. Zhang, Y. Yan, M. Lades, "Face recognition: Eigenfaces, elastic matching, and neural nets", *Proceedings of the IEEE*, Vol. 85, No. 9, 1997, pp. 1422-1435.

Fabien Cardinaux received the Master of science degree in Computer Engineering, Computer Vision and Image processing from Université de Bourgogne (Dijon, France) in 2001. He is currently a PhD candidate at the IDIAP Research Institute (Martigny, Switzerland). His research interests include computer vision, biometrics, machine learning and pattern recognition.

Conrad Sanderson received the Bachelor of Engineering (Hons) degree in 1996 and the PhD degree in 2003 from Griffith University, Queensland, Australia. He has worked at the Advanced Telecommunication Research (ATR) Laboratories (Japan), IDIAP Research Institute (Switzerland), the University of Adelaide (Australia), and is presently with National ICT Australia (NICTA). His current research interests include application and theoretical areas of biometrics and machine learning.

Samy Bengio obtained his PhD in computer science from Universite de Montreal (1993), and spent three post-doctoral years at CNET, the research center of France Telecom, and INRS-Telecommunications (Montreal). He then worked as a researcher for CIRANO, an economic and financial academic research center, applying learning algorithms to finance; he was then a research director at Microcell Labs, a private research center in mobile telecommunications. Since 1999 he is with the IDIAP Research Institute, as a senior researcher in machine learning. His current interests include all theoretical and applied aspects of learning algorithms.