IDIAP RESEARCH REPORT

ON PERFORMANCE / ROBUSTNESS / COMPLEXITY TRADE-OFFS IN FACE VERIFICATION

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

IDIAP–RR 04-74

DECEMBER 2004

(a) **conradsand @ ieee.org**; Electrical and Electronic Engineering, University of Adelaide, SA 5005, Australia.
(b) **cardinau @ idiap.ch**
(d) **bengio @ idiap.ch**

IDIAP Research Report 04-74

On Performance / Robustness / Complexity Trade-Offs in Face Verification

Conrad Sanderson Fabien Cardinaux Samy Bengio

December 2004

Abstract. In much of the literature devoted to face recognition, experiments are performed with controlled images (e.g. manual face localization, controlled lighting, background and pose); however, a practical recognition system has to be robust to more challenging conditions. In this paper we first evaluate, on the relatively difficult BANCA database, the performance, robustness and complexity of Gaussian Mixture Model (GMM), 1D- and pseudo-2D Hidden Markov Model (HMM) based systems, using both manual and automatic face localization. We also propose to extend the GMM approach through the use of local features with embedded positional information, increasing performance without sacrificing its low complexity. Experiments show that good performance on manually located faces is not necessarily indicative of good performance on automatically located faces (which are imperfectly located). The deciding factor is shown to be the degree of constraints placed on spatial relations between face parts. Methods which utilize rigid constraints have poor robustness compared to methods which have relaxed constraints. Furthermore, we show that while the pseudo-2D HMM approach has the best overall performance, classification time on current hardware makes it impractical. The best trade-off in terms of complexity, robustness and discrimination performance is achieved by the extended GMM approach.

1 Introduction

Recognizing people by biometrics (such as fingerprints, faces, speech and iris patterns) has applications in surveillance, forensics, transaction authentication, and various forms of access control, such as border checkpoints and access to digital information [14, 16, 23].

In this paper we exclusively focus on identity verification (a two-class recognition task) 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 verified [16]. Various techniques have been proposed for face classification; some examples are systems based on Principal Component Analysis (PCA) feature extraction [24], modular PCA [17], Elastic Graph Matching (EGM) [6, 12], and Support Vector Machines [20]. Examples specific to statistical models include one-dimensional Hidden Markov Models (1D HMMs) [21], pseudo-2D HMMs [7] and Gaussian Mixture Models (GMMs) [3, 22] (which can be considered as a simplified version of HMMs). A recent review of related literature can be found in [11].

GMM and HMM 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) [10] or DCTmod2 [22] is usually applied to each block, resulting in a set of feature vectors. In an analogous manner, 2D Gabor wavelets [13] can also be used.

In HMM based approaches, the spatial relations between major face features (such as the eyes and nose) is kept (although not rigidly); in the GMM approach the spatial relations are effectively lost (as each block is treated independently). As the loss of spatial information may degrade discrimination performance, in this paper we first propose to restore some of the relations 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.

Face recognition results in the literature are often presented assuming manual face localization (e.g. see [7, 15, 21]); in only relatively few publications performance evaluation is found while using automatic face localization (e.g. [3, 20]). While assuming manual (i.e. 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 it is necessary to automatically locate the face. 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).

We show that the performance of the overall face verification 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. It is shown that robustness depends on the degree of constraints placed on spatial relations between face parts.

We also 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 a verification. 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 verify an identity 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 2. Section 3 briefly describes the employed automatic face localization and feature extraction methods, while Section 4 provides a brief description of the BANCA database and its experiment protocols. Section 5 is devoted to experiments involving manual and automatic face localization; the complexity of the models is also discussed. Conclusions and future areas of research are given in Section 6.

2 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 $\lambda_{generic}$. 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:

$$\Lambda(X) = \log P(X|\lambda_C) - \log P(X|\lambda_{generic}) \tag{1}$$

where $P(X|\lambda_C)$ is the likelihood of the claim coming from the true claimant and $P(X|\lambda_{generic})$ is used as an approximation of the likelihood of the claim coming from an impostor. The verification decision is then reached as follows: given a threshold τ, the claim is accepted when $\Lambda(X) \geq \tau$ and rejected when $\Lambda(X) < \tau$.

The parameters for the generic model are found using the Expectation Maximization (EM) algorithm [5] using data from all training faces. The parameters (λ_C) for each client are found by adapting the generic model using a form of Maximum *a Posteriori* (MAP) adaptation [9, 19].

2.1 Gaussian Mixture Model

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

$$P(X|\lambda) = \prod_{t=1}^{T} P(\mathbf{x}_t|\lambda) = \prod_{t=1}^{T} \sum_{k=1}^{N_G} w_k \, \mathcal{N}(\mathbf{x}_t|\mu_k, \mathbf{\Sigma}_k) \tag{2}$$

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

2.1.1 Embedding Positional Information

If each feature vector in the set X describes a different part of the face, then a classifier based purely on GMMs effectively loses the spatial relations between face parts. We conjecture that the relations carry discriminatory information, and propose to restore a degree of the relations in the GMM approach 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. Furthermore, since the extension is done in the feature domain, the relative simplicity of the GMM approach is retained. Formally, an extended feature vector for position (a, b) is obtained with:

$$\mathbf{x}_{(a,b)}^{\texttt{extended}} = \left[\begin{array}{ccc} \left(\mathbf{x}_{(a,b)}^{\texttt{original}}\right)^T & a & b \end{array} \right]^T \tag{3}$$

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.

2.2 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 [21] or left-right HMM in the context of speech recognition [18]. 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

$$b_j(\mathbf{x}_t) = p(\mathbf{x}_t | q_t = S_j) \tag{4}$$

   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 $\lambda = (A, B)$. If we let Q be a state sequence $q_1, q_2, \cdots, q_T$, then the likelihood of an observation sequence X is:

$$P(X|\lambda) = \sum_{\forall\ Q} P(X, Q|\lambda) \ = \sum_{\forall\ Q} \prod_{t=1}^{T} b_{q_t}(\mathbf{x}_t) \prod_{t=2}^{T} a_{q_{t-1}, q_t} \tag{5}$$

The calculation of this likelihood according to the direct definition in Eqn. (5) involves an exponential number of computations. In practice the Forward-Backward procedure is used [18]; it is mathematically equivalent, but considerably more efficient.

Compared to the GMM approach described in Section 2.1, 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 not rigid, 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).

2.3 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 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 [7, 21] and the training process is described in detail in [15]. 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). 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).

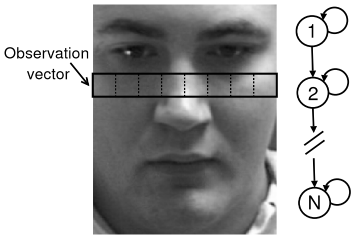

Figure 1: 1D HMM topology.

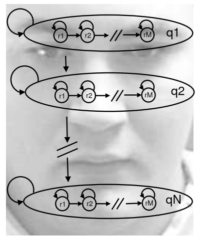

Figure 2: P2D HMM: the emission distributions of the vertical HMM are estimated by horizontal HMMs. q_i represent the states of the main HMM and r_j represent the embedded HMMs states.

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.

3 Face Localization and Feature Extraction

For automatic face localization experiments, we use the face detector recently proposed by Fröba and Ermst [8] (which is partly based on Viola and Jones' approach [25]). Eye positions are inferred from the location and scale of the bounding box enveloping the face. If no face is detected in a given image, we perform the verification using, if available, other images supporting the claim. If all given images are deemed not to contain a face, the claim is considered to have come from an impostor.

Based on the 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 [22]. 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. Each block is decomposed in terms of 2D Discrete Cosine Transform (DCT) basis functions [10]. A feature vector for each block located at row a and column b is then constructed as

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

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; the deltas are computed using DCT coefficients extracted from neighbouring blocks.

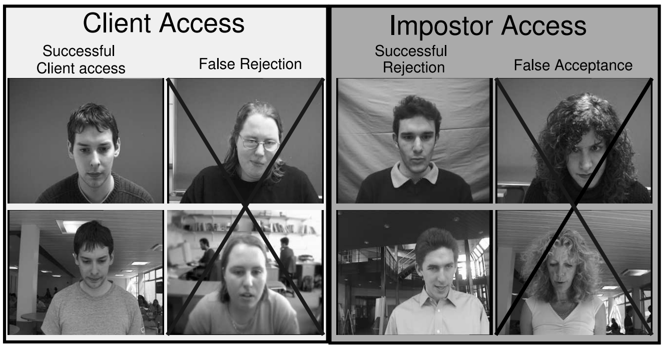

Figure 3: Example of correct and incorrect verifications on the BANCA database. Top row contains training images (from the controlled condition) while the bottom row contains test images from degraded and adverse conditions.

In this study we use M=15 (based on [22]), resulting in an 18 dimensional feature vector for each block.

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.

4 BANCA Database and Experiment Protocols

The multi-lingual BANCA database [1] was designed to evaluate multi-modal identity verification 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 data 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. 3 for an example of the differences.

We believe that the most realistic cases are when we train the system in controlled conditions and test it in different conditions. Hence in our experiments we use the Matched Controlled (Mc), Unmatched Degraded (Ud), Unmatched Adverse (Ua) and the Pooled test (P) experiment protocols, which are described in detail in [1].

To increase the number of subjects, we merged the English and French corpora, resulting in a total of 104 subjects. As per the protocol specifications, the resulting population was then equally divided into *validation* and *test* sets. Subjects in the validation set are used to optimize the verification system (e.g. to find the optimum number of gaussians and the decision threshold), while subjects from the test set are used for final performance evaluation.

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

System	Number of states		Gaussians per state	Total gaussians
	main HMM	embedded HMM		
GMM	-	-	-	512
GMMext	-	-	-	1024
1D HMM	32	-	1	32
P2D HMM	16	4	64	4096

Table 1: Optimum parameters for systems based on GMM (standard features), GMMext (extended features), 1D HMM and P2D HMM.

defined as:

$$\text{FAR} = \frac{\text{number of FAs}}{\text{number of impostor accesses}} \tag{6}$$

$$\text{FRR} = \frac{\text{number of FRs}}{\text{number of true claimant accesses}} \tag{7}$$

To aid the interpretation of performance, the two error measures are often combined using the Half Total Error Rate (HTER), defined as [2]:

$$\text{HTER} = (\text{FAR} + \text{FRR})/2 \tag{8}$$

A special case of the HTER, known as the Equal Error Rate (EER), occurs when the system is adjusted (e.g. via tuning a threshold) so that FAR=FRR on a particular data set.

5 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 decision threshold τ; the parameters were chosen to minimize the EER. The final performance of each model was then found on the test set.

Table 1 shows the optimum 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 the P2D HMM approach utilizes the largest number of gaussians, followed by the GMMext approach. The 1D HMM approach uses the least number of gaussians.

For comparison purposes, we also evaluated 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 amount of client specific training data, 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	Protocol			
	Mc	Ud	Ua	P
PCA *man.*	9.5	20.9	20.8	18.4
PCA *auto*	22.4	29.7	33.7	29.0
GMM *man.*	8.9	17.3	20.9	17.0
GMM *auto*	9.5	21.0	24.8	19.5
GMMext *man.*	8.5	17.6	20.8	16.4
GMMext *auto*	8.5	18.4	22.5	19.1
1D HMM *man.*	6.9	16.3	17.0	14.7
1D HMM *auto*	13.8	25.9	23.4	21.7
P2D HMM *man.*	4.6	15.3	13.1	13.5
P2D HMM *auto*	6.5	15.9	14.7	14.7

Table 2: HTER performance for manual face localization (*man.* suffix) and automatic face localization (*auto* suffix).

system has been used in [23]. Feature vectors with 160 dimensions were found to provide optimal performance on the validation set.

In Section 5.1 we present the results for manual face localization, Section 5.2 contains results for imperfect and automatic face localization and finally in Section 5.3 we compare the complexity of the local feature approaches.

5.1 Manual Face Localization

Table 2 shows the results in terms of HTER for manual face localization. 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 somewhat better performance. The P2D HMM approach obtains the best overall performance.

5.2 Imperfect and Automatic Face Localization

Prior to using the automatic face locator, we first study how each system is affected by 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 5.1 (i.e. trained with manually localized faces). The eye positions were artificially perturbed using:

$$eye_x = eye_x^{\mathtt{gt}} + \xi_x \tag{9}$$

$$eye_y = eye_y^{\mathtt{gt}} + \xi_y \tag{10}$$

where $eye_x^{\mathtt{gt}}$ and $eye_y^{\mathtt{gt}}$ are the ground-truth (original) co-ordinates for an eye. ξ is a random variable and follows 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. When $V = 1$, the largest translation (in one axis) will tend to be about half of the distance between the eyes.

Results in Fig. 4 show that GMM, GMMext and P2D HMM based systems are quite robust to imperfect face localization. In contrast, the PCA and 1D HMM systems are significantly more sensitive, with their discrimination performance rapidly decreasing as the error 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 2 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. When using 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.

5.3 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 verification. 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 3 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 verify 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 verification 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 times presented are indicative.

As in our implementation of MAP training only the means are adapted, the number of client specific parameters is the sum of the parameters for the means (dependent on the dimensionality of feature vectors). The other parameters (e.g. weights, covariance matrices and transition probabilities) 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).

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 verification are quite important.

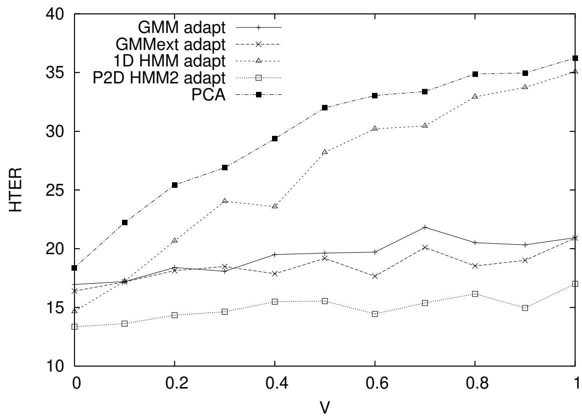

Figure 4: HTER for an increasing amount of error in eye locations.

Model type	GMM	GMMext	1D HMM	P2D HMM
number of client specific parameters	9,216	20,480	4,032	73,728
world model training time	470s (355s)	679s (546s)	192s (14s)	7967s (7789s)
client model training time	2s (1s)	3s (1.5s)	3s (2s)	251s (250s)
time for verification of one claim (5 images)	1.12s (0.24s)	1.28s (0.40s)	1.31s (0.22s)	19.89s (18.80s)

Table 3: **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).

There shouldn't be a long delay between a user enrolling in the system and being able to use the system; most importantly, the verification time should not be cumbersome, in order to aid the adoption of the verification system. The GMM, GMMext and 1D HMM approaches have short training and verification 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 verification. The P2D HMM approach has a considerably higher training and verification time, at approximately 4 minutes for training each client model and 20 seconds for a verification. With current computing resources, this verification time can be considered as being too long for practical deployment purposes. Hence in practical terms, the GMMext approach obtains the best trade-off in terms of verification time, robustness and discrimination performance.

6 Conclusions and Future Work

The findings of this paper can be summarized as follows:

- 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 locator cannot guarantee perfect face localization, this indicates that any new technique must be designed from the ground up to handle imperfectly located faces.
- Ordering the systems based on their degree of spatial constraints (loose to rigid) results in: GMM, GMMext, P2D HMM, 1D HMM and finally PCA. A system based on 1D HMMs has rigid constraints along one axis, while a system based on holistic PCA features has rigid constraints along both axes.
- 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 automatically 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.

We envisage that the performance of the extended GMM approach can be increased. Currently the degree of influence of positional information during modeling is not controlled; higher performance might be attained if more weight is placed on this information. Furthermore, the P2D HMM approach could also gain from using feature vectors with embedded positional information, as in effect more spatial constraints (though still not rigid) would be placed on each face. The main limitation of the P2D HMM system is its time requirements. The system could be deliberately detuned (e.g. by reducing the number of gaussians in each state) in order to reduce its computational complexity, and hence reduce the time taken to perform a verification. This will probably come at the cost of a loss in discrimination performance, though the extent of this loss remains to be seen. Use of embedded positional information in the feature vectors may mitigate this possible performance loss.

Acknowledgements

The authors thank J. Mariéthoz and S. Searle 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 [4].

References

[1] 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.

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

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

[15] A. Nefian and 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] C. Sanderson, K.K. Paliwal, "Identity Verification Using Speech and Face Information", *Digital Signal Processing*, Vol. 14, No. 5, 2004, pp. 449-480.

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

[25] P. Viola and 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.