IDIAP RESEARCH REPORT

Face Processing & Frontal Face Verification

Conrad Sanderson (*)

IDIAP–RR 03-20

April 2003
(Minor revision: February 2004)

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

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

(*) conradsand @ ieee.org

IDIAP Research Report 03-20

FACE PROCESSING & FRONTAL FACE VERIFICATION

Conrad Sanderson

APRIL 2003
(MINOR REVISION: FEBRUARY 2004)

Abstract. In this report we first review important publications in the field of face recognition; geometric features, templates, Principal Component Analysis (PCA), pseudo-2D Hidden Markov Models, Elastic Graph Matching, as well as other points are covered; important issues, such as the effects of an illumination direction change and the use of different face areas, are also covered. A new feature set (termed *DCT-mod2*) is then proposed; the feature set utilizes polynomial coefficients derived from 2D Discrete Cosine Transform (DCT) coefficients obtained from horizontally & vertically neighbouring blocks. Face authentication results on the VidTIMIT database suggest that the proposed feature set is superior (in terms of robustness to illumination changes and discrimination ability) to features extracted using four popular methods: PCA, PCA with histogram equalization pre-processing, 2D DCT and 2D Gabor wavelets; the results also suggest that histogram equalization pre-processing increases the error rate and offers no help against illumination changes. Moreover, the proposed feature set is over 80 times faster to compute than features based on 2D Gabor wavelets. Further experiments on the Weizmann Database also show that the proposed approach is more robust than 2D Gabor wavelets and 2D DCT coefficients.

Contents

List of Figures

List of Tables

Mathematical Notation

$\vec{x}$	a column vector
$\vec{x}^T$	vector transpose of $\vec{x}$
x_i	i-th element of vector $\vec{x}$, e.g., $\vec{x}^T = [\ x_1\ x_2\ ...\ x_D\ ]$, or, $\vec{x}^T = [\ x_i\ ]_{i=1}^{D}$
$\vec{x}_i$	i-th vector in a set
$\{\vec{x}_i\}_{i=1}^{N_V}$	set of N_V vectors
A^T	matrix transpose of A
A^{-1}	inverse of matrix A
$\|A\|$	determinant of matrix A
$\mathbf{\Sigma}$	covariance matrix
λ	parameter set (e.g., parameters of a GMM)

Acronyms

ATM	Automatic Teller Machine
BMS	Background Model Set
DCT	Discrete Cosine Transform
EER	Equal Error Rate
EGM	Elastic Graph Matching
EM	Expectation Maximization
FA	False Acceptance
FA%	False Acceptance rate
FR	False Rejection
FR%	False Rejection rate
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HTER	Half Total Error Rate, defined as HTER = $\frac{1}{2}$(FA% + FR%)
LDA	Linear Discriminant Analysis
PCA	Principal Component Analysis
UBM	Universal Background Model

1 Introduction

The field of face recognition can be divided into two areas: face identification and face verification (also known as authentication). A face verification system verifies the claimed identity based on images (or a video sequence) of the claimant's face; this is in contrast to an identification system, which attempts to find the identity of a given person out of a pool of N people.

Verification systems pervade our every day life; for example, Automatic Teller Machines (ATMs) employ simple identity verification where the user is asked to enter their password (known only to the user), after inserting their ATM card; if the password matches the one prescribed to the card, the user is allowed access to their bank account. However, the verification system such as the one used in the ATM only verifies the validity of the combination of a certain possession (in this case, the ATM card) and certain knowledge (the password). The ATM card can be lost or stolen, and the password can be compromised (e.g. somebody looks over your shoulder while you're keying it in). In order to address this issue, biometric verification methods have emerged where the password can be either replaced by, or used in addition to, biometrics such as the person's speech, face image or fingerprints. More information about the field of biometrics can be found in the following papers: [4, 16, 58, 69].

The rest of the report is organized as follows. In Section 2 we provide a concise review of previous approaches to automatic face recognition. In Section 3 we provide mathematical details for facial feature extraction techniques based on Principal Component Analysis (PCA), 2D Gabor wavelets and 2D Discrete Cosine Transform (DCT); we then propose several new feature extraction methods which build from the 2D DCT. The performance of the traditional and proposed feature extraction techniques is compared in Section 4, using artificial and real-life illumination direction changes.

This report is an extended version of [54]. The reader may also be interested in related works: [7, 8, 56, 57].

2 Summary of Past Face Recognition Approaches

This section presents a concise review of previous approaches to automatic face recognition. It goes into detail with the most important and/or popular approaches; the reader is also directed to recent survey articles [10, 23, 70].

Generally speaking, a full face recognition system can be thought of as being comprised of three stages:

1. Face localization and segmentation
2. Normalization
3. The actual face identification/verification, which can be further subdivided into:
   (a) Feature extraction
   (b) Classification

From here on we shall assume that the face has been located, or that images given to the system contain only one face, set against a uniform background. In other words, we shall concentrate on the last stage (3). Some recent approaches to face location and segmentation are presented in [24, 47, 52, 68]. The second stage (normalization) usually involves an affine transformation [22] (to correct for size and rotation), but it can also involve an illumination normalization (however, illumination normalization may not be necessary if the feature extraction method is robust against varying illumination).

There are many approaches to face recognition - ranging from the Principal Component Analysis (PCA) approach (also known as eigenfaces) [40, 63], Elastic Graph Matching (EGM) [14, 33], Artificial Neural Networks [34, 64], to pseudo-2D Hidden Markov Models (HMM) [17, 53]. All these systems differ in terms of the feature extraction procedure and/or the classification technique used. These systems, and many others, are described in the sections below.

2.1 Geometric Features vs Templates

Brunelli and Poggio [5] compared the performance of a system utilizing automatically extracted geometric features and a classifier based on the squared Mahalanobis distance [15] (similar to a single-Gaussian GMM; see Section 4.1) against a system using a template matching strategy. In the former system, the geometrical features included:

- eyebrow thickness and vertical position at the eye center position
- coarse description of the left eyebrow's arches
- vertical position and width of the nose
- vertical position of the mouth as well as the width and height
- set of radii describing the chin shape
- face width at nose position
- face width halfway between nose tip and eyes

In the latter system, four sub-images (automatically extracted from the frontal face image), representing the eye, nose, mouth and face area (from eyebrows downward), were used by a classifier based on normalized cross correlation with a set of template images. In both systems, the size of the face image was first normalized. Brunelli and Poggio found that the template matching approach obtained superior identification performance and was significantly simpler than the geometric feature based approach. Moreover, they have also found that the face areas can be sorted by discrimination ability as follows: eyes, nose and mouth; they note that this ordering is consistent with human ability of identifying familiar people from a single facial characteristic.

2.2 Principal Component Analysis (eigenfaces) and Related Techniques

Inspired by the work of Kirby and Sirovich [28], Turk and Pentland [63] proposed the use of Principal Component Analysis (PCA) [43] as a holistic feature extraction method for use in face recognition.

Given a face image matrix F of size $Y \times X$, all the columns of F are concatenated to form a column vector $\vec{f}$ of dimensionality YX. A D-dimensional feature vector, $\vec{x}$, is then obtained by:

$$\vec{x} = U^T(\vec{f} - \vec{f}_\mu) \tag{1}$$

where matrix U contains D eigenvectors (with largest corresponding eigenvalues) of the training data covariance matrix, and $\vec{f}_\mu$ is the mean of training face vectors. The eigenvectors are referred to as "eigenfaces" (see Section 3.1.1 for full derivation).

As $\vec{x}$ is in effect a dimensionality reduced version of $\vec{f}$, the above PCA based feature extraction technique is sensitive to translation, rotation, scaling as well as changes in illumination. Thus prior to feature extraction, the face image must be normalized (e.g., the location of the eyes must be the same for each person and any illumination changes must be compensated).

On a database of 16 people and using a Euclidean distance based classifier, Turk and Pentland obtained 100% identification when using face images obtained in non-challenging conditions. However, the performance decreased when there was a change in the lighting conditions, head size or head orientation.

Moghaddam and Pentland [40] modified the PCA based face recognition system to use separate face areas (i.e., eyes, nose and mouth) in a similar manner to Brunelli and Poggio [5]. By disregarding the mouth area, Moghaddam and Pentland showed that the system is less affected by expression and other changes to the face (such as a beard). Moreover, an improvement in identification rate was achieved by combining the holistic PCA system with the modular PCA system. In a separate development in the same paper, the holistic PCA system was modified to use face images processed by an edge detector, resulting in a drop in performance. The edge detector had the effect of removing most of the texture information from the face, indicating that such information is useful in recognition.

Belhumeur et al. [3] investigated the use of Linear Discriminant Analysis (LDA) as a feature extraction technique robust to changes in illumination direction. The training paradigm involved the use of face images with varying illumination. Experiments on two small databases (the largest having 16 persons) showed that the LDA based approach is significantly more robust than the PCA approach; the experiments also showed that the PCA approach can be made more robust by disregarding the first three eigenfaces, indicating that they are primarily due to lighting variation. However, when the experiment setup was modified to use training images with constant illumination and testing images with varying illumination, LDA derived features were shown to be still affected, although significantly less than PCA derived features.

2.3 Pseudo-2D Hidden Markov Model (HMM) Based Techniques

Samaria [53] extended 1D HMMs (popular in speech recognition [25, 46]) to pseudo-2D HMMs for use in face recognition. A pseudo-2D HMM for each person consists of a pseudo-2D lattice of states, each describing a distribution of feature vectors belonging to a particular area of the face. Samaria used a multivariate Gaussian [see Eqn.(27)] as a model of the distribution of feature vectors for each state. During testing, an optimal alignment of the states was found for a given image (i.e., the likelihood of each pseudo-2D HMM was maximized). Person identification was achieved by selecting the pseudo-2D HMM which obtained the highest likelihood.

Due to the alignment stage, the pseudo-2D HMM approach is inherently robust to translation, indicating that the face normalization stage need not be as accurate as for the PCA based approach.

Samaria showed that on a 40 person database the pseudo-2D HMM approach outperformed a system comprised of a nearest neighbour classifier and PCA derived feature vectors. The best pseudo-2D HMM approach used 25 states and 96 dimensional feature vectors. The face image was analyzed on a block by block basis; the grey level pixel values inside each block were arranged into a feature vector. For the PCA based approach the number of eigenfaces was varied from 5 to 199; the performance generally leveled off when 40 eigenfaces were used.

In related work, Nefian and Hayes [44] proposed to use 2D Discrete Cosine Transform (2D DCT) coefficients [22] rather than the grey level pixel values. Only the coefficients which contained most of the energy were used in forming a feature vector. The same identification rate was achieved as for grey level pixel values, but the classification time was reduced by an order of magnitude.

Eickeler et al. [17] extended the pseudo-2D HMM approach to use 2D DCT coefficients directly from JPEG compressed images [65, 66]; moreover, they have also shown that utilizing a three-Gaussian GMM to model for the distribution of feature vectors for each state outperforms a multivariate Gaussian model (i.e., a single-Gaussian GMM).

2.4 Elastic Graph Matching (EGM) Based Techniques

Lades et al. [33] proposed to use Elastic Graph Matching (EGM) for face recognition. Each face is represented by a set of feature vectors positioned on the nodes of a coarse 2D grid placed on the face. Each feature vector is comprised of a set of responses of biologically inspired 2D Gabor wavelets [35], differing in orientation and scale (see Section 3.1.2 for more information).

Comparing two faces is accomplished by matching and adapting the grid of a test image (T) to the grid of a reference image (R), where both grids have the same number of nodes; moreover, the test grid has initially the same structure as the reference grid. The elasticity of the test grid allows accommodation of face distortions (e.g., due to expression change) and to a lesser extent, changes in the view point. The quality of a match is evaluated using a distance function:

$$d(T, R) = \sum_{i=1}^{N_N} d_f(T_i, R_i) + \xi \sum_{i=1}^{N_N} d_s(T_i, R_i) \qquad (2)$$

where N_N is the number of nodes, $d_f(T_i, R_i)$ describes the difference between feature vectors representing the i-th node of the test and reference grids, while $d_s(T_i, R_i)$ describes the difference between the spatial distances of node T_i to its neighbouring nodes and the spatial distances of node R_i to its neighbouring nodes. The coefficient ξ controls the stiffness of the test grid, with large values penalizing distortion of the test grid with respect to the reference grid (thus $d_s(\cdot, \cdot)$ is used to preserve the topology between the test and reference grids).

$d(T, R)$ is minimized via translation of the test grid and perturbation of the locations of its nodes. Lades et al. proposed an approximate solution to the minimization problem, comprised of two consecutive stages. First, an approximate match is found by translating the test grid while keeping it rigid [this corresponds to the limit $\xi \to \infty$ in Eqn. (2)]. In the second stage, ξ is set to a finite value to permit small grid distortions. Each node of the test grid is visited in a random order and its location is perturbed randomly. Each stage is deemed to have reached convergence once a predefined number of trials has failed to reduce $d(T, R)$. Once convergence is reached, the value of $d(T, R)$ is used for recognition purposes. Lades et al. reported encouraging identification results where test faces contained expression changes and small rotations.

Duc et al. [14] extended the EGM approach to include node specific weighting of the contribution of each Gabor wavelet response to the measure of the difference between feature vectors. On a database which had mainly expression changes, the extended system provided lower verification error rates than the standard system. Moreover, Duc et al. showed that the extended system still outperformed the standard system even if the second stage of minimization of $d(T, R)$ is omitted (i.e., the test grid is kept rigid).

Kotropoulos et al. [31] used the outputs of multiscale morphological dilation and erosion operations [22] to yield a feature vector for each node. Compared to feature vectors based on responses of Gabor wavelets, the advantage of the morphological operation approach is that it is significantly faster due to its relative simplicity and lack of floating point arithmetic operations. Comparative verification results in [61] show that the morphological operation based approach has slightly lower error rates than the standard approach based on Gabor wavelets.

2.5 Other Approaches

Matas et al. [38] proposed a face verification method based on a robust form of correlation. A search for the optimum correlation is performed in the space of all valid geometric and photometric transformations of the test image to obtain the best match with the reference image. The geometric transformation includes translation, rotation and scaling, while the photometric transformation corrects the mean of pixel intensity across the face. The quality of the match between a transformed test image and a reference image is evaluated using a sum of pixel differences, subject to a constraint: if the pixel difference is above a predefined threshold, it is ignored. This constraint is utilized in order to discount face regions which are subject to relatively large change (such as hair style and expression). The search technique involves the random selection of transformation parameters; each transformation is accepted only if the matching score is increased. To speed up the search, a randomly selected subset of pixels is used instead of the entire image. Verification results on a database which had mainly expression changes show a minor improvement over Duc's extended EGM approach (described in Section 2.4).

Lawrence et al. [34] proposed the use of a hybrid neural-network approach to face recognition. The system combined local image sampling, a self-organizing map (SOM) [30] and a convolutional neural network. On a database of 40 people, the proposed approach obtained an identification error rate of 3.8%, compared to 10.5% obtained using a system comprised of the PCA based feature extractor (described Section 2.2) and a nearest neighbour classifier. By replacing the features obtained using local image sampling and the SOM with PCA derived features it was shown the improvement in performance can be largely attributed to the convolutional neural network (i.e., the classifier).

2.6 Important Issues

Zhang et al. [70] compared the performance of the EGM approach with a system comprised of the PCA based feature extractor and a nearest neighbour classifier. Results on a combined database of 100 people showed that the PCA based system was more robust to scale and rotation variations, while the EGM approach was more robust to position, illumination and expression variations. Zhang et al. contributed the robustness to illumination changes to the use of Gabor features, while the robustness to position and expression variations was contributed to the deformable matching stage.

Kotropoulos et al. [32] showed that while morphologically derived feature vectors are more sensitive to illumination changes than Gabor wavelet derived features, they are less sensitive to face size variations. They proposed a heuristic size and illumination normalization technique, which, on a small database containing face images collected in real life conditions, was shown to significantly improve the performance of a EGM based system which utilized the morphologically derived feature vectors. Strangely, no comparative results were reported for Gabor wavelet derived feature vectors.

Adini et al. [1] studied the suitability of several image processing techniques for reducing the effects of an illumination direction change (where one side of the face was brighter than the other). Various configurations of the following techniques were considered: filtering with 2D Gabor-like filters [35], edge maps, first & second derivatives and log transformations [22]. Several classifiers, based on pixel differences between two processed images, were also evaluated; all of the classifiers produced similar identification results. On a database comprised of 25 subjects, Adini et al. found that none of the processing techniques were sufficient to completely overcome the effects of the illumination direction change; most techniques obtained an identification rate of less than 50%. However, when using unprocessed images, the identification rate was 0%. Adini et al. showed that the 2D Gabor-like filter which emphasized the differences along the vertical axis (e.g., the eyebrows and the eyes) obtained the best results. This is not surprising, considering that the illumination direction change produced the greatest pixel intensity changes along the horizontal axis. Moreover, results obtained using the vertical orientation were mostly independent of the scale of the filter; at other orientations, the size of the filter greatly affected the identification rate. These results indicate that the optimum orientation and scale of the 2D Gabor-like filter is dependent on the direction of the illumination change.

Belhumeur et al. [3] found that the recognition rate is significantly higher when using *full faces* (that is, containing the hair and the outline of the face) than when using *closely cropped* faces (that is, containing only the eyebrows, eyes, nose and mouth), indicating that the overall shape of the face is an important feature. However, Belhumeur et al. conjectured that the recognition rate would drop significantly for the *full faces* if the background or hairstyles were varied; moreover, it may be even lower than for *closely cropped* faces. Chen et al. [11] quantitatively proved that the influence of the *closely cropped* area on the recognition process is much smaller than that of the outside area (i.e., the hair and the outline of the face). By using synthetic *full face* images, where the hair and face outline of one person was combined with the *closely cropped* area from another person, Chen et al. successfully confused a PCA based face recognition system. Along with the results of Moghaddam and Pentland [40] (see Section 2.2), these results indicate that for a statistics based face recognition system, the area containing the eyebrows, eyes and the nose is the most useful. The mouth area needs to be disregarded as it is mostly affected by expression changes and beards.

3 Feature Extraction for Face Verification

From the review in Section 2 it is evident that PCA derived features, and to a lesser extent, 2D Gabor wavelet derived features, are affected by an illumination direction change. As will be shown, 2D DCT based features are also sensitive to changes in the illumination direction. In this section we introduce four new feature sets, which are significantly less affected by an illumination direction change: *DCT-delta*, *DCT-mod*, *DCT-mod-delta* and *DCT-mod2*. We will show that the *DCT-mod2* method, which utilizes polynomial coefficients derived from 2D DCT coefficients of spatially neighbouring blocks, is the most suitable. We then compare the robustness and performance of the *DCT-mod2* method against two popular feature extraction techniques, eigenfaces (PCA) and 2D Gabor wavelets, in addition to the standard 2D DCT approach.

The rest of this section is organized as follows. In Section 3.1, we review the PCA, 2D Gabor wavelet and 2D DCT feature extraction methods, and describe the proposed feature extraction methods.

To keep consistency with traditional matrix notation, pixel locations (and image sizes) are described using the row(s) first, followed by the column(s).

3.1 Feature Extraction Techniques

3.1.1 Eigenfaces (PCA)

Given a face image matrix[1] F of size $Y \times X$, we construct a vector representation by concatenating all the columns of F to form a column vector $\vec{f}$ of dimensionality YX. Given a set of training vectors $\{\vec{f}_i\}_{i=1}^{N_P}$ for all persons, we define the mean of the training set as $\vec{f}_\mu$. A new set of mean subtracted vectors is formed using:

$$\vec{g}_i = \vec{f}_i - \vec{f}_\mu, \quad i = 1, 2, ..., N_P \tag{3}$$

The mean subtracted training set is represented as matrix $G = [\ \vec{g}_1\ \vec{g}_2\ ...\ \vec{g}_{N_P}]$. The covariance matrix is calculated using:

$$C = GG^T \tag{4}$$

Due to the size of C, calculation of the eigenvectors of C can be computationally infeasible. However, if the number of training vectors (N_P) is less than their dimensionality (YX), there will be only $N_P - 1$ meaningful eigenvectors. Turk and Pentland [63] exploit this fact to determine the eigenvectors using an alternative method, summarized as follows. Let us denote the eigenvectors of matrix G^TG as $\vec{v}_j$ with corresponding eigenvalues λ_j:

$$G^TG\vec{v}_j = \lambda_j \vec{v}_j \tag{5}$$

Pre-multiplying both sides by G gives us:

$$GG^TG\vec{v}_j = \lambda_j G\vec{v}_j \tag{6}$$

Letting $\vec{u}_j = G\vec{v}_j$ and substituting for C from Eqn. (4):

$$C\vec{u}_j = \lambda_j \vec{u}_j \tag{7}$$

Hence the eigenvectors of C can be found by pre-multiplying the eigenvectors of G^TG by G. To achieve dimensionality reduction, let us construct matrix $U = [\ \vec{u}_1\ \vec{u}_2\ ...\ \vec{u}_D\ ]$, containing D eigenvectors of C with

[1] The face images used in our experiments have 56 rows (Y) and 64 columns (X).

largest corresponding eigenvalues. Here, $D < N_P$. A feature vector $\vec{x}$ of dimensionality D is then derived from a face vector $\vec{f}$ using:

$$\vec{x} = U^T(\vec{f} - \vec{f_\mu}) \tag{8}$$

i.e., face vector $\vec{f}$ decomposed in terms of D eigenvectors, known as "eigenfaces".

3.1.2 2D Gabor Wavelets

The biologically inspired family of 2D Gabor wavelets is defined as follows [35]:

$$\Psi(y,x,\omega,\theta) = \frac{\omega}{\kappa\sqrt{2\pi}}\psi_A(y,x,\omega,\theta)\left[\psi_B(y,x,\omega,\theta) - \exp\left\{-\frac{\kappa^2}{2}\right\}\right] \tag{9}$$

where

$$\psi_A(y,x,\omega,\theta) = \exp\left\{-\frac{\omega^2}{8\kappa^2}\left[4(y\sin\theta + x\cos\theta)^2 + (y\cos\theta - x\sin\theta)^2\right]\right\} \tag{10}$$

and

$$\psi_B(y,x,\omega,\theta) = \exp\left\{i(\omega y\sin\theta + \omega x\cos\theta)\right\} \tag{11}$$

Here ω is the radial frequency in radians per unit length and θ is the wavelet orientation in radians. Each wavelet is centered at point $(y,x) = (0,0)$. The family is made up of wavelets for N_ω radial frequencies, each with N_θ orientations. The radial frequencies are spaced in octave steps and cover a range from $\omega_{min} > 0$ to $\omega_{max} < \pi$, where 2π represents the Nyquist frequency. Typically $\kappa \approx \pi$ so that each wavelet has a frequency bandwidth of one octave [35].

Feature extraction is done as follows. A coarse rectangular grid is placed over given face image F. At each node of the grid, the inner product of F with each member of the family is computed:

$$P_{j,k} = \int_y \int_x \Psi(y_0 - y, x_0 - x, \omega_j, \theta_k) F(y,x)\, dx\, dy \tag{12}$$

for $j = 1,2,...,N_\omega$ and $k = 1,2,...,N_\theta$. Here, the node is located at (y_0, x_0). An $N_\omega N_\theta$-dimensional feature vector[2] for location (y_0, x_0), is then constructed using the modulus of each inner product [33]:

$$\vec{x} = \left[\ |P_{1,1}|\ \ |P_{1,2}|\ \cdots\ |P_{1,N_\omega}|\ \cdots\ |P_{2,1}|\ \ |P_{2,2}|\ \cdots\ |P_{2,N_\omega}|\ \cdots\ |P_{N_\theta,N_\omega}|\ \right]^T \tag{13}$$

Thus if there are N_G nodes in the grid, we extract N_G feature vectors from one image.

3.1.3 2D Discrete Cosine Transform

Here the given face image is analyzed on a block by block basis. Given an image block $f(y,x)$, where $y,x = 0,1,...,N-1$ (typically $N = 8$), we decompose it in terms of orthogonal 2D DCT basis functions (see Figure 1). The result is an $N \times N$ matrix $C(v,u)$ containing 2D DCT coefficients:

$$C(v,u) = \alpha(v)\alpha(u)\sum_{y=0}^{N-1}\sum_{x=0}^{N-1} f(y,x)\beta(y,x,v,u) \quad \text{for } v,u = 0,1,2,...,N-1 \tag{14}$$

where

$$\alpha(v) = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } v = 0 \\ \sqrt{\frac{2}{N}} & \text{for } v = 1,2,...,N-1 \end{cases} \tag{15}$$

[2]Typically, $N_\omega = 3$ and $N_\theta = 6$, resulting in an 18 dimensional vector.

and

$$\beta(y,x,v,u) = \cos\left[\frac{(2y+1)v\pi}{2N}\right]\cos\left[\frac{(2x+1)u\pi}{2N}\right] \tag{16}$$

The coefficients are ordered according to a zig-zag pattern, reflecting the amount of information stored [22] (see Figure 2). For a block located at (b,a), the baseline 2D DCT feature vector is composed of:

$$\vec{x} = \left[\; c_0^{(b,a)} \;\; c_1^{(b,a)} \;\; \ldots \;\; c_{M-1}^{(b,a)} \;\right]^T \tag{17}$$

where $c_n^{(b,a)}$ denotes the n-th 2D DCT coefficient and M is the number of retained coefficients[3]. To ensure adequate representation of the image, each block overlaps its horizontally and vertically neighbouring blocks by 50% [17]. Thus for an image which has Y rows and X columns, there are $N_D = (2\frac{Y}{N} - 1) \times (2\frac{X}{N} - 1)$ blocks[4].

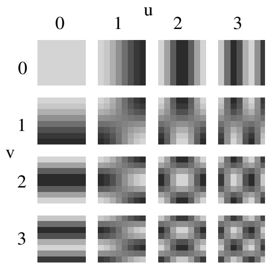

Figure 1: Several 2D DCT basis functions for N=8. Lighter colours represent larger values.

v \ u	**0**	**1**	**2**	**3**
0	0	1	5	6
1	2	4	7	12
2	3	8	11	13
3	9	10	14	15

Figure 2: Ordering of 2D DCT coefficients $C(v,u)$ for N=4.

3.1.4 Proposed DCT-delta

In speech based systems, features based on polynomial coefficients (also known as deltas), representing transitional spectral information, have been successfully used to reduce the effects of background noise and channel mismatch [60].

For images, we define the n-th *horizontal* delta coefficient for block located at (b,a) as a modified 1st order orthogonal polynomial coefficient [26, 60]:

$$\Delta^h c_n^{(b,a)} = \frac{\sum_{k=-K}^{K} k h_k \, c_n^{(b,a+k)}}{\sum_{k=-K}^{K} h_k k^2} \tag{18}$$

Similarly, we define the n-th *vertical* delta coefficient as:

$$\Delta^v c_n^{(b,a)} = \frac{\sum_{k=-K}^{K} k h_k \, c_n^{(b+k,a)}}{\sum_{k=-K}^{K} h_k k^2} \tag{19}$$

where h is a $2K+1$ dimensional symmetric window vector. In this report we shall use $K = 1$ and a rectangular window (thus $\vec{h} = [\, 1.0 \; 1.0 \; 1.0 \,]^T$).

Let us assume that we have three horizontally consecutive blocks X, Y and Z. Each block is composed of two components: facial information and additive noise; e.g., $X = I_X + I_N$. Moreover, let us also suppose that

[3] In our experiments, $M = 15$.

[4] Thus for a 56×64 (rows $\times$ columns) image, there are 195 2D DCT feature vectors.

all of the blocks are corrupted with the same noise (a reasonable assumption if the blocks are small and close or overlapping). To find the deltas for block Y, we apply Eqn. (18) to obtain (ignoring the denominator):

$$\begin{aligned} \Delta^h Y &= -X + Z && (20) \\ &= -(I_X + I_N) + (I_Z + I_N) && (21) \\ &= I_Z - I_X && (22) \end{aligned}$$

i.e., the noise component is removed.

By combining the horizontal and vertical delta coefficients an overall delta feature vector is formed. Hence, given that we extract M 2D DCT coefficients from each block, the delta vector is $2M$ dimensional. We shall term this feature extraction method as *DCT-delta*.

DCT-delta feature extraction for a given block is only possible when the block has vertical and horizontal neighbours. Thus processing an image which has Y rows and X columns and using a 50% block overlap results in $N_{D2} = (2\frac{Y}{N} - 3) \times (2\frac{X}{N} - 3)$ *DCT-delta* feature vectors[5].

3.1.5 Proposed DCT-mod, DCT-mod2 and DCT-mod-delta

By inspecting Eqns. (14) and (16), it is evident that the 0-th 2D DCT coefficient will reflect the average pixel value (or the DC level) inside each block and hence will be the most affected by any illumination change. Moreover, by inspecting Figure 1 it is evident that the first and second coefficients represent the average horizontal and vertical pixel intensity change, respectively. As such, they will also be significantly affected by any illumination change. Hence we shall study three additional feature extraction approaches (in all cases we assume the baseline 2D DCT feature vector is M dimensional):

1. Discard the first three coefficients from the baseline 2D DCT feature vector. We shall term this *modified* feature extraction method as *DCT-mod*.
2. Discard the first three coefficients from the baseline 2D DCT feature vector and concatenate the resulting vector with the corresponding *DCT-delta* feature vector. We shall refer to this method as *DCT-mod-delta*.
3. Replace the first three coefficients with their horizontal and vertical deltas, and form a feature vector representing a given block as follows:

$$\vec{x} = \left[\; \Delta^h c_0 \; \Delta^v c_0 \; \Delta^h c_1 \; \Delta^v c_1 \; \Delta^h c_2 \; \Delta^v c_2 \;\; c_3 \; c_4 \; ... \; c_{M-1} \; \right]^T \quad (23)$$

   where the (b, a) superscript was omitted for clarity. Let us term this modified approach as *DCT-mod2*.

Thus each *DCT-mod-delta* and *DCT-mod2* feature vector represents transitional spatial information as well as local texture information.

As for *DCT-delta*, *DCT-mod-delta* and *DCT-mod2* feature extraction for a given block is only possible when the block has vertical and horizontal neighbours. Thus processing an image which has Y rows and X columns and using a 50% block overlap results in $N_{D2} = (2\frac{Y}{N} - 3) \times (2\frac{X}{N} - 3)$ *DCT-mod-delta* or *DCT-mod2* feature vectors[6].

4 Evaluation

This section is dedicated to evaluating the performance of the traditional and proposed feature extraction techniques described in the last section. In Section 4.1 we describe a Gaussian Mixture Model (GMM) based

[5] Thus for a 56×64 image, there are 143 *DCT-delta* feature vectors.
[6] Thus for a 56×64 image, there are 143 *DCT-mod-delta* or *DCT-mod2* feature vectors.

classifier which shall be used as the basis for experiments. In Section 4.2 we describe the VidTIMIT database which is employed in experiments utilizing an artificial illumination direction change; the experiments are described in Section 4.3 and the results are discussed in Section 4.4. Section 4.5 is devoted to experiments on the Weizmann Database [1] which has more realistic illumination direction changes.

4.1 GMM Based Classifier

Given a claim for person C's identity and a set of feature vectors $X = \{\vec{x}_i\}_{i=1}^{N_V}$ supporting the claim, the average log likelihood of the claimant being the true claimant is calculated using:

$$\mathcal{L}(X|\lambda_C) = \frac{1}{N_V}\sum_{i=1}^{N_V}\log p(\vec{x}_i|\lambda_C) \tag{24}$$

$$\text{where} \quad p(\vec{x}|\lambda) = \sum_{j=1}^{N_G} m_j\,\mathcal{N}(\vec{x};\vec{\mu}_j,\mathbf{\Sigma}_j) \tag{25}$$

$$\lambda = \{m_j,\vec{\mu}_j,\mathbf{\Sigma}_j\}_{j=1}^{N_G} \tag{26}$$

Here, $\mathcal{N}(\vec{x};\vec{\mu},\mathbf{\Sigma})$ is a D-dimensional Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$:

$$\mathcal{N}(\vec{x};\vec{\mu},\mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\exp\left[\frac{-1}{2}(\vec{x}-\vec{\mu})^T\mathbf{\Sigma}^{-1}(\vec{x}-\vec{\mu})\right] \tag{27}$$

λ_C is the parameter set for person C, N_G is the number of Gaussians and m_j is the weight for Gaussian j (with constraints $\sum_{j=1}^{N_G} m_j = 1$ and $\forall\, j : m_j \geq 0$).

Given the average log likelihood of the claimant being an impostor, $\mathcal{L}(X|\lambda_{\overline{C}})$, an opinion on the claim is found using:

$$\Lambda(X) = \mathcal{L}(X|\lambda_C) - \mathcal{L}(X|\lambda_{\overline{C}}) \tag{28}$$

The verification decision is reached as follows: given a threshold t, the claim is accepted when $\Lambda(X) \geq t$ and rejected when $\Lambda(X) < t$.

4.1.1 Model Training & Impostor Likelihood

Given a set of training vectors, $X = \{\vec{x}_i\}_{i=1}^{N_V}$ (which may come from several images), the GMM parameters (λ) for each client model are found by the Expectation Maximization (EM) algorithm [12, 42, 15].

The likelihood of the claimant being an impostor can be found via the use of a composite model[7], comprised of several GMMs for other clients. The client models in such a composite are referred to as background models [48] or cohort models [20]. Given N_B background models, the impostor likelihood is found using:

$$\mathcal{L}(X|\lambda_{\overline{C}}) = \log\left[\frac{1}{N_B}\sum_{b=1}^{N_B} p(X|\lambda_b)\right] \tag{29}$$

The background model set contains models which are the "closest" as well as the "farthest" from the client model [48]. While it may intuitively seem that only the "close" models are required (which represent the expected impostors), this would leave the system vulnerable to impostors which are very different from the client. This is demonstrated by inspecting Eqn. (28) where both terms would contain similar likelihoods, leading to an unreliable opinion on the claim.

In this report we have utilized the method described in [48] to select the background models for each client.

[7] It must be noted that the Universal Background Model [49] can also be used to find $\mathcal{L}(X|\lambda_{\overline{C}})$.

4.2 VidTIMIT Audio-Visual Database

The VidTIMIT database [55], is comprised of video and corresponding audio recordings of 43 people (19 female and 24 male), reciting short sentences. It was recorded in 3 sessions, with a mean delay of 7 days between Session 1 and 2, and 6 days between Session 2 and 3. There are 10 sentences per person; the first six sentences are assigned to Session 1; the next two sentences are assigned to Session 2 with the remaining two to Session 3. The mean duration of each sentence is 4.25 seconds, or approximately 106 video frames.

The video of each person is stored as a sequence of high quality JPEG images with a resolution of 384×512 pixels. The corresponding audio is stored as a mono, 16 bit, 32 kHz WAV file.

4.3 Experiments

Before feature extraction can occur, the face must first be located. Furthermore, to account for varying distances to the camera, a geometrical normalization must be performed. Here we treat the problem of face location and normalization as separate from feature extraction.

To find the face, we use template matching with several prototype faces[8] of varying dimensions. Using the distance between the eyes as a size measure, an affine transformation is used [22] to adjust the size of the image, resulting in the distance between the eyes to be the same for each person. Finally a 56×64 pixel face window, $w(y, x)$, containing the eyes and the nose (the most invariant face area to changes in the expression and hair style) is extracted from the image.

For PCA, the dimensionality of the face window is reduced to 40 (choice based on the works by Kirby and Sirovich [28], Samaria [53] and Belhumeur et al. [3]).

For 2D DCT and 2D DCT derived methods, each block is 8×8 pixels. Moreover, each block overlaps with horizontally and vertically adjacent blocks by 50%.

For Gabor wavelet features, we heed the choice of Duc et al. [14] with $N_\omega = 3$, $N_\theta = 6$, $\omega_1 = \frac{\pi}{2}$, $\omega_2 = \frac{\pi}{4}$, $\omega_3 = \frac{\pi}{8}$ and $\theta_k = \frac{\pi(k-1)}{N_\theta}$ (where $k = 1, 2, ..., N_\theta$). Hence the dimensionality of the Gabor feature vectors is 18. The location of the wavelet centers was chosen to be as close as possible to the centers of the blocks used in *DCT-mod2* feature extraction.

In our experiments, we use a sequence of images (video) from the VidTIMIT database for person verification. If the sequence has N_I images, then $N_V = N_I$ for PCA derived features, $N_V = N_I N_G$ for Gabor features, $N_V = N_I N_D$ for 2D DCT and *DCT-mod* features and $N_V = N_I N_{D2}$ for *DCT-delta*, *DCT-mod-delta* and *DCT-mod2* features. To reduce the computational burden during modeling and testing, every second video frame was used. For each feature extraction method, 8-Gaussian[9] client models (GMMs) were generated from features extracted from face windows in Session 1. Sessions 2 and 3 were used for testing. Thus for each person an average of 318 frames were used for training and 212 for testing.

Ignoring any edges created by shadows, the main effect of an illumination direction change is that one part of the face is brighter than the rest[10]. Taking this into account, an illumination direction change was introduced to face windows extracted from Sessions 2 and 3; to simulate more illumination on the left side of the face and

[8] A “mother” prototype face was constructed by averaging manually extracted and size normalized faces from all people in the VidTIMIT database; prototype faces of various sizes were constructed by applying an affine transform to the “mother” prototype face.

[9] Choice based on preliminary experiments.

[10] As evidenced by the images presented in [31], which were obtained under real-life conditions.

less on the right, a new face window $v(y,x)$ is created by transforming $w(y,x)$ using[11]:

$$\begin{aligned} v(y,x) &= w(y,x) + mx + \delta \\ \text{for: } y &= 0, 1, \cdots, N_Y - 1 \\ x &= 0, 1, \cdots, N_X - 1 \\ \text{where: } m &= \frac{-\delta}{(N_X - 1)/2} \\ \delta &= \text{illumination delta (in pixels)} \end{aligned} \quad (30)$$

Example face windows for various δ are shown in Figure 3. It must be noted that this model of illumination direction change is artificial and restrictive as it does not cover all the effects possible in real life (shadows[12], etc.), but it is useful for providing suggestive results[13].

Figure 3: Examples of varying light illumination; left: $\delta = 0$ (no change); middle: $\delta = 40$; right: $\delta = 80$

To find the performance, Sessions 2 and 3 were used for obtaining example opinions of known impostor and true claims. Four utterances, each from 8 fixed persons (4 male and 4 female), were used for simulating impostor accesses against the remaining 35 persons. As in [48], 10 background person models were used for the impostor likelihood calculation. For each of the remaining 35 persons, their four utterances were used separately as true claims. In total there were 1120 impostor and 140 true claims. The decision threshold was then set so the *a posteriori* performance is as close as possible to the Equal Error Rate (EER) [i.e. where the False Acceptance rate (FA%) is equal to the False Rejection rate (FR%)]. This protocol is described in more detail in [55].

In the first experiment, we found the performance of the 2D DCT approach on face windows with $\delta = 0$ (i.e., no illumination change) while varying the dimensionality of the feature vectors. The results are presented in Figure 4. As can be observed, the performance improves immensely as the number of dimensions is increased from 1 to 3. Increasing the dimensionality from 15 to 21 provides only a relatively small improvement, while significantly increasing the amount of computation time required to generate the models. Based on this we have chosen 15 as the dimensionality of baseline 2D DCT feature vectors; hence the dimensionality of *DCT-delta* feature vectors is 30, *DCT-mod* is 12, *DCT-mod-delta* is 42 and *DCT-mod2* is 18.

In the second experiment we compared the performance of 2D DCT and all of the proposed techniques for increasing δ. Results are shown in Figure 5.

In the third experiment we compared the performance of PCA, PCA with histogram equalization pre-processing[14], DCT, Gabor and *DCT-mod2* features for varying δ. Results are presented in Figure 6.

In the fourth experiment, we have evaluated the effects of varying block overlap used during *DCT-mod2* feature extraction (in all other experiments, the overlap was fixed at 50%). Varying the overlap has two effects: the first is that as overlap is increased the spatial area used to derive one feature vector is decreased; the second

[11]Please note that many authors (for example, [19, 67]) describe light changes as a multiplicative effect on image brightness. In our experiments we have treated the face image simply as an information source. The transformation described by Eqn. (30) is in effect an empirical information transformation method; it has been designed to transform the face information to approximate the face images presented in [31].

[12]However, the face images presented in [3] show that only extreme illumination direction conditions produce significant shadows, where even humans have trouble recognizing faces.

[13]See also Section 4.5 for experiments on the Weizmann Database [1].

[14]Histogram equalization [9, 22] is often used in an attempt to reduce the effects of varying illumination conditions [29, 41].

Method	Time (msec)
PCA	11
DCT	6
Gabor	675
DCT-mod2	8

Table 1: Average time taken per face window (results obtained using Pentium III 500 MHz, Linux 2.2.18)

effect is that the number of feature vectors extracted from an image grows in an exponential manner as the overlap is increased. Results are shown in Figure 7.

Computational burden is an important factor in practical applications, where the amount of required memory and speed of the processor have direct bearing on the final cost. Hence in the final experiment we compared the average time taken to process one face window by PCA, DCT, Gabor and *DCT-mod2* feature extraction techniques. It must be noted that apart from having the transformation data pre-calculated (e.g., β 2D DCT basis functions), no thorough hand optimization of the code was done. Nevertheless, we feel that this experiment provides figures which are at least indicative. Results are listed in Table 1.

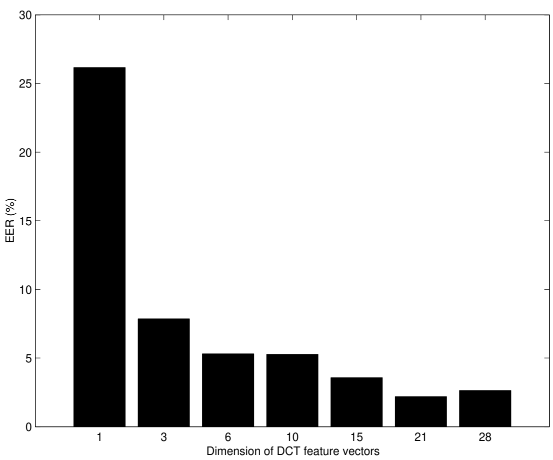

Figure 4: Performance for varying dimensionality of 2D DCT feature vectors

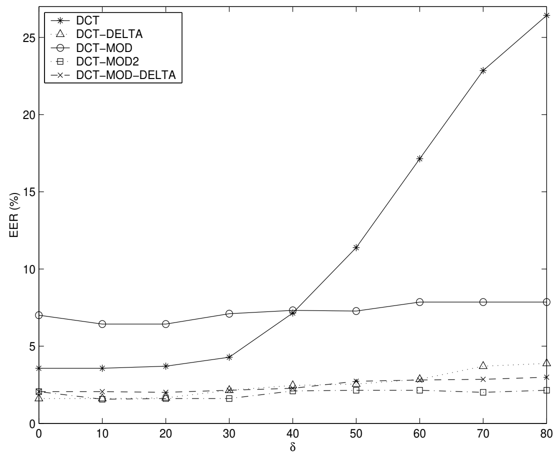

Figure 5: Performance of 2D DCT and proposed feature sets

4.4 Discussion

As can be observed in Figure 4, the first three 2D DCT coefficients contain a significant amount of person dependent information; thus ignoring them (as in *DCT-mod*) implies a reduction in performance. This is verified in Figure 5 where the *DCT-mod* features have worse performance than 2D DCT features when there is little or no illumination direction change ($\delta \leq 30$). We can also see that the performance of DCT features is fairly stable for small illumination direction changes but rapidly degrades for $\delta \geq 40$ (in contrast to *DCT-mod* features which have a relatively static performance).

The remaining feature sets (*DCT-delta*, *DCT-mod-delta* and *DCT-mod2*) do not have the performance penalty associated with the *DCT-mod* feature set. Moreover, all of them have similarly better performance than 2D DCT features; we conjecture that the increase in performance can be attributed to the effectively larger spatial area used when obtaining the features. *DCT-mod2* edges out *DCT-delta* and *DCT-mod-delta* in terms of stability for large illumination direction changes ($\delta \geq 50$). Additionally, the dimensionality of *DCT-mod2* (18)

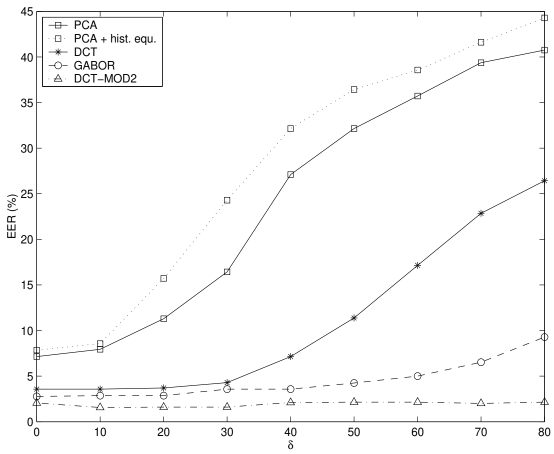

Figure 6: Performance of PCA, PCA with histogram equalization pre-processing, DCT, Gabor and *DCT-mod2* feature sets

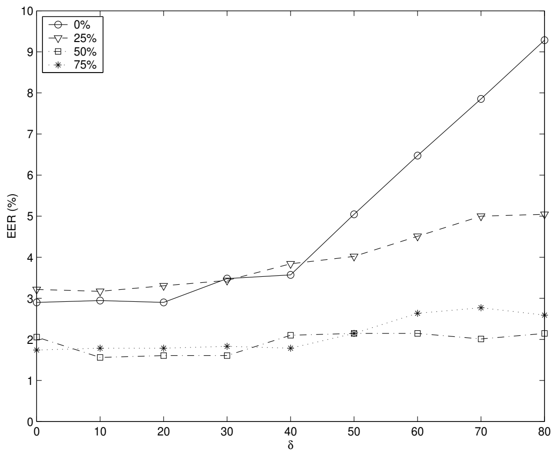

Figure 7: Performance of *DCT-mod2* feature set for varying overlap

is lower than *DCT-delta* (30) and *DCT-mod-delta* (42).

The results suggest that delta features make the system more robust as well as improve performance; they also suggest that it is only necessary to use deltas of coefficients representing the average pixel intensity and low frequency features (i.e. the 0-th, first and second 2D DCT coefficients) while keeping the remaining DCT coefficients unchanged; hence out of the four proposed feature extraction techniques, the *DCT-mod2* approach is the most suitable.

Using 0% or 25% block overlap in *DCT-mod2* feature extraction (Fig. 7) results in a performance degradation as δ is increased, implying that the assumption that the blocks are corrupted with the same noise has been violated (see Section 3.1.4). Increasing the overlap from 50% to 75% had little effect on the performance at the expense of extracting significantly more feature vectors.

By comparing the performance of PCA, PCA with histogram equalization pre-processing, 2D DCT, 2D Gabor and *DCT-mod2* feature sets (Figure 6), it can be seen that the *DCT-mod2* approach is the most immune to illumination direction changes (the performance is virtually flat for varying δ). The performance of PCA derived features rapidly degrades as δ increases, while the performance of 2D Gabor features is stable for $\delta \leq 40$ and then gently deteriorates as δ increases. We can also see that use of histogram equalization as pre-processing for PCA increases the error rate in all cases, and most notably offers no help against illumination changes. The results thus suggest that we can order the feature sets, based on their robustness and performance, as follows: *DCT-mod2*, 2D Gabor, 2D DCT, PCA, and lastly, PCA with histogram equalization pre-processing.

From Table 1 we can see that 2D Gabor features are the most computationally expensive to calculate, taking about 84 times longer than *DCT-mod2* features. This is due to the size of the 2D Gabor wavelets as well as the need to compute both real and imaginary inner products. Compared to 2D Gabor features, PCA, 2D DCT and *DCT-mod2* features take a relatively similar amount of time to process one face window.

It must be noted that when using the GMM classifier in conjunction with the 2D Gabor, 2D DCT or *DCT-mod2* features, the spatial relation between major face features (e.g., eyes and nose) is lost. However, excellent performance is still obtained, implying that the use of more complex classifiers which preserve spatial relation, such as a pseudo-2D HMM and elastic graph matching, is not necessary. Moreover, due to the loss of the spatial relations[15], the GMM classifier theoretically has some inbuilt robustness to translation (which may be caused by inaccurate face localization).

[15]See also Appendix A.

Method	Illumination direction		
	uniform	left	right
DCT	3.49	48.15	48.15
Gabor	0.36	33.34	33.34
DCT-mod2	0	25.93	22.65

Table 2: Results on the Weizmann Database, quoted in terms of HTER

It must also be noted that using the introduced illumination change, the center portion of the face (column wise) is largely unaffected; the size of the portion decreases as δ increases. In the PCA approach one feature vector describes the entire face, hence any change to the face would alter the features obtained. This is in contrast to the other approaches (2D Gabor, 2D DCT and *DCT-mod2*), where one feature vector describes only a small part of the face. Thus a significant percentage (dependent on δ) of the feature vectors is largely unchanged, automatically leading to a degree of robustness.

4.5 Experiments on the Weizmann Database

The experiments described in Section 4.3 utilize an artificial illumination direction change. In this section we shall compare the performance of 2D DCT, 2D Gabor wavelet and *DCT-mod2* feature sets (see Section 3) on the Weizmann Database [1], which has more realistic illumination direction changes.

It must be noted that the database is rather small, as it is comprised of images of 27 people; moreover, for the direct frontal view, there is only one image per person with uniform illumination (the training image) and two test images where the illumination is either from the left or right; all three images were taken in the same session. As such, the database is not suited for verification experiments, but some suggestive results can still be obtained.

The experimental setup is similar to that described in Section 4.3. However, due to the small amount of training data, an alternative GMM training strategy is used. Rather than training the client models directly using the EM algorithm, each model is derived from a Universal Background Model (UBM) by means of maximum *a posteriori* (MAP) adaptation [21, 49]. The UBM is trained via the EM algorithm using pooled training data from all clients. Moreover, due to the small number of persons in the database, the UBM is also used to calculate the impostor likelihood (rather than using a set of background models). A detailed description of this training and testing strategy is presented in [49].

Since PCA based feature extraction produces one feature vector per image (see Section 3.1.1), there is insufficient training data to reliably train the client models. Thus PCA based feature extraction is not evaluated here.

For each illumination type, the client's own training image was used to simulate a true claim. Images from all other people were used to simulate impostor claims. In total, for each illumination type, there were 27 true claims and 702 impostor claims. The *a posteriori* decision threshold was set to obtain performance as close as possible to EER. Results are presented in Table 2, in terms of Half Total Error Rate (HTER), defined as HTER = $\frac{1}{2}$(FA%+FR%).

As can be observed, no method is immune to the changes in the illumination direction. However *DCT-mod2* features are the least affected, followed by Gabor features and lastly DCT features.

5 Conclusion

In this report we first reviewed important publications in the field of face recognition. Geometric features, templates, Principal Component Analysis (PCA), pseudo-2D Hidden Markov Models (HMM), Elastic Graph Matching (EGM), as well as other points were covered. Important issues, such as the effects of an illumination direction change and the use of different face areas, were also covered. Several new feature extraction techniques were proposed; out of the proposed methods, the *DCT-mod2* technique, which utilizes polynomial coefficients derived from 2D DCT coefficients of spatially neighbouring blocks, is the most suitable. Face verification results on the VidTIMIT database suggest that the *DCT-mod2* feature set is superior (in terms of robustness to illumination direction changes and discrimination ability) to features extracted using four popular methods: eigenfaces (PCA), PCA with histogram equalization pre-processing, 2D DCT and 2D Gabor wavelets; the results also suggest that histogram equalization pre-processing increases the error rate and offers no help against illumination changes. Moreover, the *DCT-mod2* feature set is over 80 times faster to compute than features based on Gabor wavelets. Further experiments on the Weizmann Database have also shown that the *DCT-mod2* approach is more robust than 2D Gabor wavelets and 2D DCT coefficients.

6 Acknowledgments

The author thanks Professor Kuldip K. Paliwal for his suggestions and fruitful discussions. The initial version of this work was written while the author was a student at Griffith University; revision was performed at IDIAP, with thanks to support by the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

A Face Areas Modeled by the GMM

A typical example of the face areas modeled by each Gaussian (in an 8-Gaussian GMM) is shown in Figure 8, where *DCT-mod2* feature extraction was used. Images from a video sequence of the face were used to train the model. The selected areas represent the center blocks used in *DCT-mod2* feature extraction (see Section 3.1.5). Some overlap between the areas for different Gaussians is present since a 50% block overlap was used.

As can be observed, the type of area modeled by each Gaussian is generally guided by the degree of smoothness of the area; this leads to automatic selection of physically meaningful areas, such as the eyes and the nose. This is expected, since the EM algorithm used to train each GMM is in effect a probabilistic clustering procedure [15], where similar features are represented by each Gaussian.

Figure 9 shows a typical example of the effect of decomposing a face image in terms of a different person's model. In this case, *fdrd1*'s model was used to decompose *mbdg0*'s face image.

By comparing Figures 8 and 9 it can be seen that *fdrd1*'s model selects similar areas in *fdrd1*'s and *mbdg0*'s face images. Thus if we assume that, in a verification scenario, subject *mbdg0* claims to be subject *fdrd1*, the GMM-based face verification system, in effect, compares *fdrd1*'s eyes against *mbdg0*'s eyes.

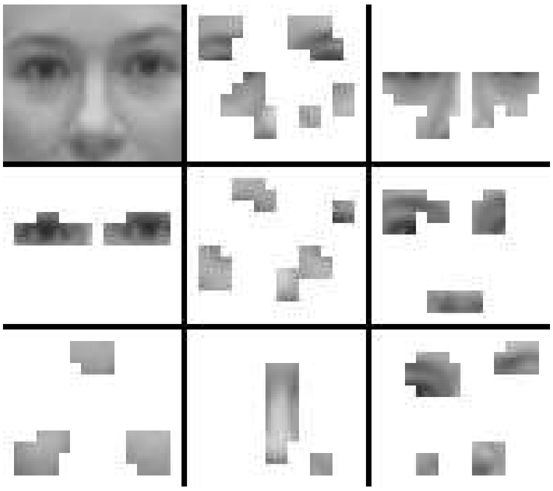

Figure 8: Typical example of 8-Gaussian GMM face modeling. Top left: original image of subject *fdrd1*; other squares: areas modeled by each Gaussian in *fdrd1*'s model (*DCT-mod2* feature extraction).

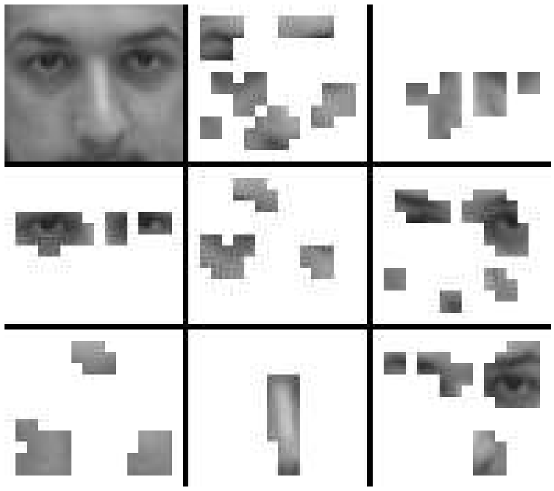

Figure 9: Top left: original image of subject *mbdg0*; other squares: areas selected by *fdrd1*'s Gaussians.

References

[1] Y. Adini, Y. Moses and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, 1997, pp. 721-732.

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

[3] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 711-720.

[4] R. M. Bolle, J. H. Connell and N. K. Ratha, "Biometric perils and patches", *Pattern Recognition*, Vol. 35, No. 12, 2002, pp. 2727-2738.

[5] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 15, No. 10, 1993, pp. 1042-1052.

[6] M. M. Buechner, "Eye of the Beholder", *Time Australia*, 27 November 2000 (No. 47), pp. 89-92.

[7] F. Cardinaux, C. Sanderson and S. Marcel, "Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS", *Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guildford, 2003, pp. 911-920.

[8] F. Cardinaux, C. Sanderson and S. Bengio, "Face Verification Using Adapted Generative Models", *Proc. Int. Conf. Automatic Face and Gesture Recognition (AFGR)*, Seoul, Korea, 2004.

[9] K. R. Castleman, *Digital Image Processing*, Prentice-Hall, USA, 1996.

[10] 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.

[11] L-F. Chen, H-Y. Liao, J-C. Lin and C-C. Han, "Why recognition in a statistics-based face recognition system should be based on the pure face portion: a probabilistic decision-based proof", *Pattern Recognition*, Vol. 34, No. 7, 2001, pp. 1393-1403.

[12] A.P. Dempster, N.M. Laird and 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.

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

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

[15] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, John Wiley & Sons, USA, 2001.

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

[17] S. Eickeler, S. Müller and G. Rigoll, "Recognition of JPEG Compressed Face Images Based on Statistical Methods", *Image and Vision Computing*, Vol. 18, No. 4, 2000, pp. 279-287.

[18] K. R. Farrell, "Text-Dependent Speaker Verification Using Data Fusion", *Proc. IEEE International Conf. Acoustics, Speech and Signal Processing*, Detroit, Michigan, 1995, Vol. 1, pp. 349-352.

[19] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, USA, 2003.

[20] S. Furui, "Recent Advances in Speaker Recognition", *Pattern Recognition Letters*, Vol. 18, No. 9, 1997, pp. 859-872.

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

[22] R. C. Gonzales and R. E. Woods, *Digital Image Processing*, Addison-Wesley, Reading, Massachusetts, 1993.

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

[24] C-C. Han, H-Y. M. Liao, G-J. Yu and L-H. Chen, "Fast face detection via morphology-based pre-processing", *Pattern Recognition*, Vol. 33, No. 10, 2000, pp. 1701-1712.

[25] X. Huang, A. Acero and H-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR, New Jersey, 2001.

[26] N. L. Johnson and F. C. Leone, *Statistics and Experimental Design in Engineering and the Physical Sciences* (Vol. 1), John Wiley & Sons, USA, 1977.

[27] K. Jonsson, J. Kittler, Y.P. Li and J. Matas, "Support Vector Machines for Face Authentication", *Image and Vision Computing*, Vol. 20, No. 5-6, 2002, pp. 369-375.

[28] M. Kirby and L. Sirovich, "Application of the Karhunen-Loève Procedure for the Characterization of Human Faces", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 12, No. 1, 1990, pp. 103-108.

[29] L. H. Koh, S. Ranganath and Y. V. Venkatesh, “An integrated automatic face detection and recognition system”, *Pattern Recognition*, Vol. 35, No. 6, 2002, pp. 1259-1273.

[30] T. Kohonen, “The self-organizing map”, *Proceedings of the IEEE*, Vol. 78, No. 9, 1990, pp. 1464-1480.

[31] C. Kotropoulos, A. Tefas and I. Pitas, “Frontal Face Authentication Using Morphological Elastic Graph Matching”, *IEEE Trans. Image Processing*, Vol. 9, No. 4, 2000, pp. 555-560.

[32] C. Kotropoulos, A. Tefas and I. Pitas, “Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions”, *Pattern Recognition*, Vol. 33, No. 12, 2000, pp. 1935-1947.

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

[34] S. Lawrence, C. L. Giles, A. C. Tsoi and A. D. Back, “Face Recognition: A Convolutional Neural-Network Approach”, *IEEE Trans. Neural Networks*, Vol. 8, No. 1, 1997, pp. 98-113.

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

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

[37] J. Markowitz, “Speech systems work together in harmony”, *Biometric Technology Today*, Vol. 9, No. 4, 2001, pp. 7-8.

[38] J. Matas, K. Jonsson and J. Kittler, “Fast face localisation and verification”, *Image and Vision Computing*, Vol. 17, No. 8, 1999, pp. 757-581.

[39] E. Messmer, “Pentagon lab may give biometrics needed boost”, *CNN.com* web site (http://www.cnn.com/2001/TECH/science/03/20/pentagon.biometrics.idg/index.html), 20 March 2001.

[40] B. Moghaddam and A. Pentland, “Probabilistic Visual Learning for Object Representation”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 696-710.

[41] H. Moon, and P. J. Phillips, “Computational and performance aspects of PCA-based face-recognition algorithms”, *Perception*, Vol. 30, 2001, pp. 303-321.

[42] T. K. Moon, “Expectation-maximization Algorithm”, *IEEE Signal Processing Magazine*, Vol. 13, No. 6, 1996, pp. 47-60.

[43] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*, Prentice Hall, Upper Saddle River, New Jersey, 2000.

[44] A. V. Nefian and M. H. Hayes, “Hidden Markov Models for Face Recognition”, *Proc. International Conf. on Acoustics, Speech and Signal Processing*, Seattle, 1998, Vol. 5, pp. 2721-2724.

[45] S. Pigeon and L. Vandendorpe, “Image-based multi-modal face authentication”, *Signal Processing*, Vol. 69, No. 1, 1998, pp. 59-79.

[46] L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall PTR, New Jersey, 1993.

[47] B. Raducanu, M. Graña, F.X. Albizuri and A. d'Anjou, "Face localization based on the morphological multiscale fingerprints", *Pattern Recognition Letters*, Vol. 22, No. 3-4, 2001, pp. 359-371.

[48] D. A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", *Speech Communication*, Vol. 17, No. 1-2, 1995, pp. 91-108.

[49] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, No. 1-3, 2000, pp. 19-41.

[50] A. E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang and F. K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification", *Proc. International Conf. Spoken Language Processing*, Alberta, 1992, Vol. 1, pp. 599-602.

[51] A.E. Rosenberg and S. Parthasarathy, "Speaker Background Models for Connected Digit Password Speaker Verification", *Proc. International Conf. Acoustics, Speech and Signal Processing*, Atlanta, 1996, Vol. 1, pp. 81-84.

[52] H. A. Rowley, S. Baluja and T. Kanade, "Neural Network-Based Face Detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, 1998, pp. 23-38.

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

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

[55] C. Sanderson, "The VidTIMIT Database", IDIAP Communication 02-06, Martigny, Switzerland, 2002.

[56] C. Sanderson and S. Bengio, "Robust Features for Frontal Face Authentication in Difficult Image Conditions", IDIAP Research Report 03-05, Martigny, Switzerland, 2003.

[57] C. Sanderson, S. Bengio, "Statistical Transformation Techniques for Face Verification Using Faces Rotated in Depth", IDIAP Research Report 04-04, Martigny, Switzerland, 2004.

[58] C. Sanderson and K.K. Paliwal, "On the Use of Speech and Face Information for Identity Verification", IDIAP Research Report 04-10, Martigny, Switzerland, 2004.

[59] F. Smeraldi and J. Bigün, "Retinal vision applied to facial features detection and face authentication", *Pattern Recognition Letters*, Vol. 23, No. 4, 2002, pp. 463-473.

[60] F. K. Soong and 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.

[61] A. Tefas, C. Kotropoulos and I. Pitas, "Face Authentication by Using Elastic Graph Matching and Support Vector Machines", *Proc. International Conf. Acoustics, Speech and Signal Processing*, Istanbul, Vol. 4, pp. 2409-2412.

[62] A. Tefas, C. Kotropoulos and I. Pitas, "Using Support Vector Machines to Enhance the Performance of Elastic Matching for Frontal Face Authentication", *IEEE Trans. Pattern Analysis and Machine Intelligence* Vol. 23, No. 7, 2001, pp. 735-745.

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

[64] D. Valentin, H. Abdi, A. J. O'Toole and G. W. Cottrell, "Connectionist Models of Face Processing: A Survey", *Pattern Recognition*, Vol. 27, No. 9, 1994, pp. 1209-1230.

[65] G. K. Wallace, "The JPEG Still Picture Compression Standard", *Communications of the Association for Computing Machinery*, Vol. 34, No. 4, 1991, pp. 30-44.

[66] G. K. Wallace, "The JPEG still picture compression standard", *IEEE Trans. Consumer Electronics*, Vol. 38, No. 1, 1992, pp. xviii-xxxiv.

[67] Y. Weiss, "Deriving intrinsic images from image sequences", *Proc. 8th IEEE International Conf. Computer Vision*, Vancouver, 2001.

[68] K-W. Wong, K-M. Lam and W-C. Siu, "An efficient algorithm for human face detection and facial feature extraction under different conditions", *Pattern Recognition*, Vol. 34, No. 10, 2001, pp. 1993-2004.

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

[70] J. Zhang, Y. Yan and M. Lades, "Face Recognition: Eigenface, Elastic Matching, and Neural Nets", *Proceedings of the IEEE*, Vol. 85, No. 9, 1997, pp. 1423-1435.