Robust Face Based Identity Verification

Conrad Sanderson and Kuldip K. Paliwal

Abstract— **In this paper we propose two feature extraction techniques, termed *DCT-mod* and *DCT-delta*, for use in an illumination invariant face based identity verification system. We compare the performance of the proposed techniques against two standard methods: Principal Component Analysis (PCA) and the 2-D Discrete Cosine Transform (DCT). Experiments on the VidTIMIT database support the use of the proposed techniques.**

I. Introduction

IDENTITY 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 Personal Identification Number (PIN), known only to the user, after inserting their ATM card. If the PIN matches the one prescribed to the card, the user is allowed access to their bank account. Similar verification systems are used to restrict access to rooms and buildings as well as equipment such as photocopiers.

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 PIN). The ATM card can be lost or stolen, and the PIN can be compromised (eg. somebody looks over your shoulder while you're entering the PIN). Hence new verification methods have emerged, where the PIN has either been replaced by, or used in addition to, biometrics such as the person's speech, face image or fingerprints. The use of biometrics is attractive since they cannot be lost or forgotten and vary significantly between people.

The basic operation of a face based verification system is as follows:

1. A claim for an identity is presented along with a supporting video of the person's face
2. The system extracts person-dependent information (known as feature extraction) from the face images and compares it against a model of the features from the person whose identity is being claimed. Let us refer to the result of this comparison as a client score, S_C.
3. The system also compares the information against a model of possible impostors. Let us refer to the result of this comparison as an impostor score, S_I.
4. An opinion, O, on the claim is found using $O = S_C \ / \ S_I$. A relatively high opinion indicates the person is a true claimant, while a relatively low opinion suggests the person is an impostor.
5. The opinion is thresholded to achieve the final decision to either accept or reject the claim.

The performance of a verification system is measured in terms of False Acceptance rate (FA%) and False Rejection rate (FR%), defined as:

$$\text{FA} = \frac{I_A}{I_T} \times 100\% \qquad \text{FR} = \frac{C_R}{C_T} \times 100\%$$

where I_A is the number of impostors classified as true claimants, I_T is the total number of impostor classification tests, C_R is the number of true claimants classified as impostors, and C_T is the total number of true claimant classification tests.

To quantify the performance into a single number, two measures can be used: Equal Error Rate (EER), where the system is configured to operate with FA = FR and Total Error (TE), defined as TE = FA + FR.

Systems based on face images often employ techniques such as Principal Component Analysis (PCA) for feature extraction [1]. While PCA is quite effective, it is sensitive to changes in the illumination direction [2] causing rapid degradation in verification performance.

The 2-Dimensional Discrete Cosine Transform (DCT) [3] has been also been used for facial feature extraction [4], [5]. However, as will be shown, this technique is also sensitive to changes in the illumination direction. We propose two new techniques (both based on the DCT), termed *DCT-mod* and *DCT-delta*, which are significantly less affected by the illumination direction.

The rest of the paper is organized as follows. In Section II we describe the PCA, DCT, *DCT-mod* and *DCT-delta* feature extraction techniques. In Section III we describe a Gaussian Mixture Model (GMM) classifier which shall be used as the basis for experiments. In Section IV we describe the VidTIMIT audio-visual database used in the experiments. The performance of all feature extraction techniques is compared in Section V.

II. Feature Extraction

A. Principal Component Analysis (PCA)

A face image can be represented by a matrix containing grey level pixel values. A straightforward approach of representing facial information with a feature vector is to simply concatenate all the columns of the matrix. However even for a very low resolution facial image, the resulting feature vector is of prohibitive dimensionality. For example, given a 64×56 pixel face image, the resulting feature vector has 3584 dimensions. Hence dimensionality reduction methods, such as the PCA approach [6] have been used. Facial features derived from PCA are also known as eigenfaces [1]. This is done as follows: given a facial image matrix F of size $X \times Y$, we construct a vector representation by concatenating all the columns of F to form a column vector $\vec{f}$ of dimensionality XY. Given a set of training vectors $\vec{f}_i, i = 1, 2, ..., N_P$ 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:

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

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:

$$C = GG^T \tag{2}$$

where G^T denotes transpose of G. Due to the size of C, calculation of eigenvectors of C can be computationally infeasible. Turk and Pentland show an alternative way to

determine the eigenvectors [1]: if the number of training vectors (N_P) is less than their dimensionality (XY), there will be only $N_P - 1$ meaningful eigenvectors. Let us denote the eigenvectors of matrix G^TG as $\vec{v_j}$ with corresponding eigenvalues λ_j:

$$G^TG\vec{v_j} = \lambda_j\vec{v_j} \tag{3}$$

Pre-multiplying both sides by G gives us:

$$GG^TG\vec{v_j} = \lambda_jG\vec{v_j} \tag{4}$$

Letting $\vec{u_j} = G\vec{v_j}$ and substituting for C from Equation (2):

$$C\vec{u_j} = \lambda_j\vec{u_j} \tag{5}$$

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 $U = [\vec{u_1}\vec{u_2}...\vec{u_D}]$ where $D < N_P$. A feature vector $\vec{x}$ of dimensionality D is then derived from a facial image $\vec{f}$ using:

$$\vec{x} = U^T(\vec{f} - \vec{f_\mu}) \tag{6}$$

B. 2-D Discrete Cosine Transform (DCT)

The 2-D DCT is a popular technique used in image compression (for example the JPEG standard). Here the given face image is analyzed on a block by block basis, where each block has a size of $N \times N$ pixels. For each block we extract an M-dimensional feature vector. Feature extraction is done as follows: given a block $f(x, y)$ where $x, y = 0, 1, ..., N-1$ we decompose it in terms of 2-D DCT basis functions (a graphical representation of these basis functions is shown in Fig. 1). The result is a $N \times N$ matrix $C(u, v)$ containing DCT coefficients. Formally, $C(u, v)$ is found as follows:

$$C(u,v) = \alpha(u)\alpha(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\beta(x,y,u,v) \tag{7}$$

$$\text{for} \quad u, v = 0, 1, 2, ..., N-1$$

$$\text{where} \quad \alpha(u) = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } u = 0 \\ \sqrt{\frac{2}{N}} & \text{for } u = 1, 2, ..., N-1 \end{cases} \tag{8}$$

$$\text{and} \quad \beta(x,y,u,v) = \cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \tag{9}$$

The M-dimensional feature vector is then formed by taking the first M DCT coefficients. The DCT coefficients $C(u, v)$ are ordered according to a zig-zag pattern, reflecting the amount of information stored in each coefficient [3]. As an example, the order of DCT coefficients for $N = 8$ is shown in Fig. 2.

The overall process of DCT feature extraction is shown in Fig. 3.

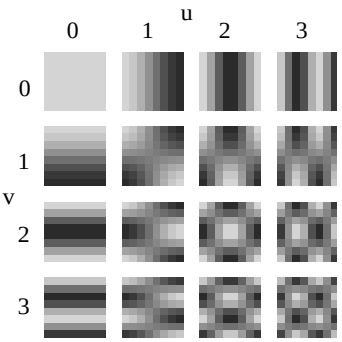

Fig. 1. Example DCT basis functions for N=8. Lighter colours represent larger values

v \ u	0	1	2	3	4	5	6	7
0	0	1	5	6	14	15	27	28
1	2	4	7	13	16	26	29	42
2	3	8	12	17	25	30	41	43
3	9	11	18	24	31	40	44	53
4	10	19	23	32	39	45	52	54
5	20	22	33	38	46	51	55	60
6	21	34	37	47	50	56	59	61
7	35	36	48	49	57	58	62	63

Fig. 2. Ordering of DCT coefficients $C(u, v)$ for $N = 8$

C. DCT-mod

By inspecting Equations (7) and (9), we can see that the 0-th DCT coefficient ($u = 0$, $v = 0$) will reflect the average pixel value (or the DC level) inside the $N \times N$ block. Hence this coefficient will be most affected by any illumination change. Moreover, by inspecting Fig. 1 it is evident that the first ($u = 1$, $v = 0$) and second coefficients ($u = 0$, $v = 1$) represent the average vertical and horizontal pixel intensity change, respectively. As such, they will also be significantly affected by any illumination change.

To reduce the effects of illumination change, we propose to ignore the 0th, 1st and 2nd coefficients when forming a feature vector from the DCT coefficients. We shall term this modified feature extraction method as *DCT-mod*.

D. DCT-delta

Delta (or regression) features have been successfully used to reduce the effects of background noise and channel mismatch in speech based systems [7]. In images, a horizontal delta feature vector can be defined as the difference between feature vectors derived from horizontally neighbouring blocks. Similarly, a vertical delta feature vector can be defined as the difference between feature vectors derived from vertically neighbouring blocks.

We propose to form an overall delta feature vector by con-

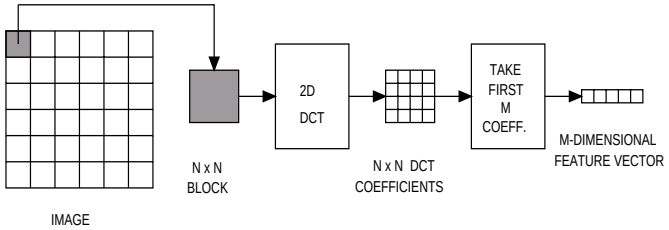

Fig. 3. Conceptual block diagram of DCT feature extraction

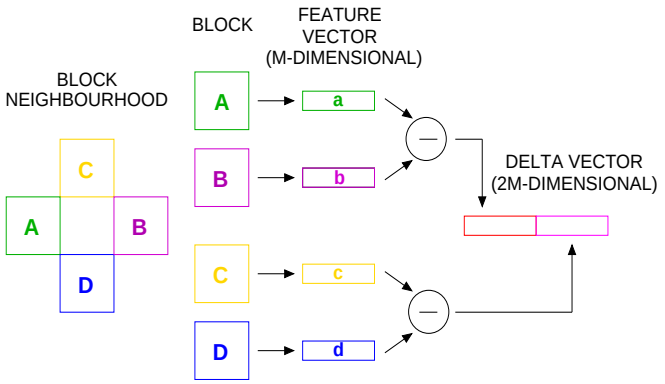

Fig. 4. Conceptual block diagram of DCT-delta feature extraction

catenating the horizontal and vertical delta feature vectors. We shall term this feature extraction method as *DCT-delta*. A block diagram of this process is shown in Fig. 4.

III. GMM Classifier

The distribution of feature vectors for each person is modeled by a Gaussian Mixture Model (GMM). Given a set of training vectors, an N_M-mixture GMM is trained using a k-means clustering algorithm followed by 10 iterations of the Expectation Maximization (EM) algorithm [8].

Given a claim for person C's identity and a set of feature vectors $X = \{\vec{x_i}, i = 1, 2, ..., N_V\}$ supporting the claim, log likelihood of the claimant being the true claimant is calculated using:

$$\log p(X|\lambda_C) = \frac{1}{N_V} \sum_{i=1}^{N_V} \log\left[p(\vec{x}_i|\lambda_C)\right] \tag{10}$$

$$\text{where} \quad p(\vec{x}|\lambda) = \sum_{i=1}^{N_M} m_i \ \mathcal{N}(\vec{x}, \vec{\mu}_m, \Sigma_m) \tag{11}$$

$$\text{and} \quad \lambda = \{m_i, \vec{\mu}_i, \Sigma_i, i = 1, 2, ..., N_M\} \tag{12}$$

Here λ_C is the model for person C. N_M is the number of mixtures, m_i is the weight for mixture i, and $\mathcal{N}(\vec{x}, \vec{\mu}, \Sigma)$ is a multi-variate Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix Σ.

Given a set of B background person models [9] (also known as cohorts) $\{\lambda_b, b = 1, 2, ..., B\}$ for person C, the log likelihood of the claimant being an impostor is found using:

$$\log p(X|\lambda_{\overline{C}}) = \log\left[\frac{1}{B} \sum_{b=1}^{B} p(X|\lambda_b)\right] \tag{13}$$

An opinion on the claim is found using the following likelihood ratio test:

$$o = \frac{p(X|\lambda_C)}{p(X|\lambda_{\overline{C}})} \tag{14}$$

In the log domain this becomes:

$$O = \log p(X|\lambda_C) - \log p(X|\lambda_{\overline{C}}) \tag{15}$$

The verification decision is reached as follows: given a threshold t, the claim is accepted when $O \geq t$; the claim is rejected when $O < t$.

A conceptual block diagram of a verification system employing the GMM is shown in Fig. 5.

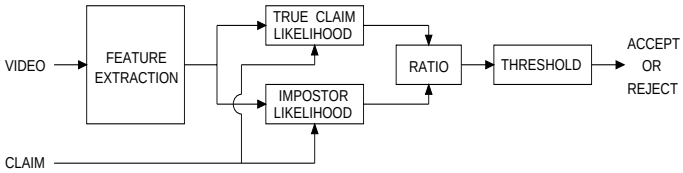

Fig. 5. Conceptual block diagram of a verification system

Fig. 6. Example images from the VidTIMIT database

IV. VidTIMIT Audio-Visual Database

With the help of many volunteers (including numerous people from within the School), we have created an audio-visual database used for experiments in person identity verification.

The database 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.

The sentences were chosen from the test section of the NTIMIT corpus [10]. 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 first two sentences for all persons are the same, with the remaining eight generally different for each person.

The recording was done in an office environment using a broadcast quality digital video camera. The video of each person is stored as a sequence of JPEG images with a resolution of 512 × 384 pixels. The corresponding audio is stored as a mono, 16-bit, 32 kHz WAV file. Several example images are presented in Figure 6. For more information, please visit *http://spl.me.gu.edu.au/vidtimit/*

V. 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.

To find the face, we use template matching with several prototype faces of varying dimensions. Using the distance between the eyes as a size measure, an affine transformation is used [3] to adjust the size of the image. This causes the distance between the eyes to be the same for each person. Finally a 64×56 pixel face window, $w(x, y)$, containing the

Fig. 7. Examples of varying light illumination; left: $\delta = 0$ (no change); middle: $\delta = 40$; right: $\delta = 80$

eyes and the nose is extracted from the image.

It must be noted that here we treat the problem of face location and normalization as separate from feature extraction.

For PCA, the dimensionality of the face window is reduced to 40. The choice of the dimensionality is based on the work by Samaria [11].

For DCT, each block is 8×8 pixels and the first 15 coefficients are retained. Moreover, each block overlapped adjacent blocks by 50%. The choice of dimensionality, block size and overlap is based on the work by Eickler [5]. Our own preliminary experiments have shown no significant improvement in performance by retaining more coefficients.

Since the dimension of DCT derived feature vectors is 15, *DCT-mod* features are 12 dimensional and *DCT-delta* features are 30 dimensional.

For each feature extraction method, 8 mixture client models (GMMs) were generated from features extracted from face windows in Session 1. Using more mixtures provided little improvement in performance and increased the model generation time significantly. Since on average there are 100 frames per sentence, every second frame was used to reduce the computational burden.

To find the decision threshold, Session 2 was used for obtaining example opinions of known impostor and true claims. Two utterances each from 8 fixed persons (4 male and 4 female) were used for simulating impostor accesses against the remaining 35 persons. 10 background person models were used for the impostor likelihood calculation. For each of the remaining 35 persons, their two utterances were used separately as true claims. In total there were 280 impostor and 70 true example claims. The decision threshold was set for EER performance.

An artificial illumination change was introduced to face windows extracted from Session 3. To simulate more illumination on the left side of the face and less on the right, a new face window $v(x, y)$ is created by transforming $w(x, y)$ using:

$$v(x, y) = w(x, y) + mx + \delta \quad (16)$$
$$\text{for} \quad x = 0, 1, ..., 64$$
$$\text{and} \quad y = 0, 1, ..., 56$$

$$\text{where} \quad m = \frac{-\delta}{64/2} \quad (17)$$
$$\text{and} \quad \delta = \text{illumination delta (in pixels)}$$

Example face windows for various δ are shown in Fig. 7.

The trained system was then tested using the face windows with varying illumination. The same setup as used for Session 2 was employed to find opinions of impostor and true claims. However, the threshold was fixed to the one found for Session 2. The results, in terms of TE, are presented in Fig. 8.

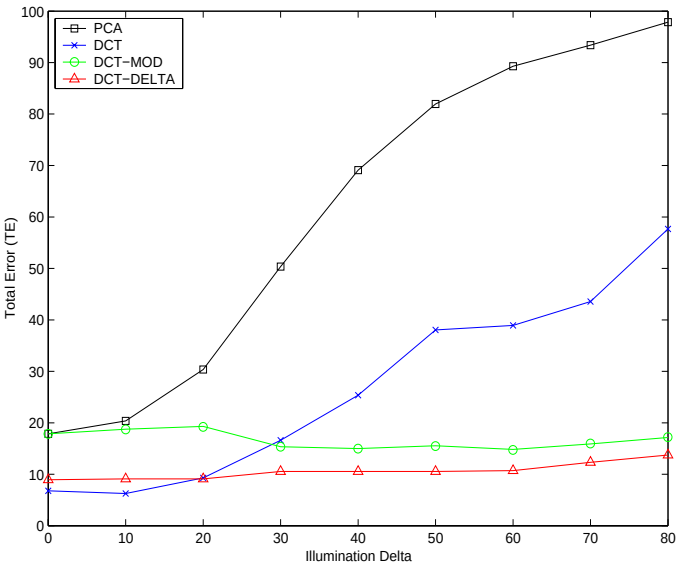

Fig. 8. Performance of various feature extraction techniques

VI. Conclusion

As it can be seen in Fig. 8, performance of both PCA and DCT approaches rapidly deteriorates as the illumination delta (δ) increases.

Using the proposed *DCT-mod* feature extraction makes the system largely invariant to illumination changes. However, this robustness comes at a price of poorer performance when there is no or little change in the illumination. These results suggest that while the first three DCT coefficients are significantly affected by illumination change, they also contain a significant amount of person dependent information.

Use of the proposed *DCT-delta* features makes the system almost invariant to changes in the illumination direction, without the large performance sacrifice present with *DCT-mod*. These results support the use of *DCT-delta* features.

References

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

[2] P. N. Belhumeur et al, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *Proc. IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, Iss. 7, 1997.

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

[4] C. Podilchuk and X. Zhang, "Face Recognition Using DCT-Based Feature Vectors", *Proc. Intern. Conf. Acoustics, Speech and Signal Processing*, 1996.

[5] S. Eickler et al, "Recognition of JPEG Compressed Face Images Based on Statistical Methods", *Image and Vision Computing 18*, 2000.

[6] C. Thierrien, "Decision estimation and classification", Wiley, 1989.

[7] F. K. Soong and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", *Proc. IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 36, No. 6, 1988.

[8] T. K. Moon, "Expectation-maximization Algorithm", *IEEE Signal Processing Magazine*, Vol. 13, Iss. 6, 1996.

[9] D. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", *Speech Communication 17*, 1995.

[10] C. Jankowski et al, "NTIMIT: A Phonetically Balanced, Continuous Speech Telephone Bandwidth Speech Database", *Proc. Intern. Conf. Acoustics, Speech and Signal Processing*, Albuquerque, 1990.

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