POLYNOMIAL FEATURES FOR ROBUST FACE AUTHENTICATION

Conrad Sanderson and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia

ABSTRACT

In this paper we introduce the *DCT-mod2* facial feature extraction technique which utilizes polynomial coefficients derived from 2-D DCT coefficients of spatially neighbouring blocks. We evaluate its robustness and performance against three popular feature sets for use in an identity verification system subject to illumination changes. Results on the multi-session VidTIMIT database suggest that the proposed feature set is the most robust, followed by (in order of robustness and performance): 2-D Gabor wavelets, 2-D DCT coefficients and PCA (eigenface) derived features. Moreover, compared to Gabor wavelets, the *DCT-mod2* feature set is over 80 times quicker to compute.

1. INTRODUCTION

A face authentication system verifies the claimed identity (a 2 class task) 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. Past research on face based systems has concentrated on the identification aspect even though the verification task has the greatest application potential [1]. This is demonstrated in security applications (eg. access control), where the claimant has good reason to cooperate with the system, as well as in forensic applications where the task is mostly evaluation of each suspect separately rather than choosing one from many persons.

While identification and verification systems share feature extraction techniques and in many cases a large part of the classifier structure, there is no guarantee that an approach used in the identification scenario would work equally well in the verification scenario.

There are many approaches to face based systems - ranging from the ubiquitous Principal Component Analysis (PCA) approach (also known as eigenfaces) [2], Dynamic Link Architecture (also known as elastic graph matching) [3], Artificial Neural Networks [4], to pseudo-2D Hidden Markov Models (HMM) [5].

These systems differ in terms of the feature extraction procedure and/or the classification technique used. For example, in [2] PCA is used for feature extraction and a nearest neighbour classifier is utilized for recognition. In [3], biologically inspired 2-D Gabor wavelets [6] are used for feature extraction, while the Dynamic Link Architecture is part of the classifier. In [5], features are derived using the 2-D Discrete Cosine Transform (DCT) and the pseudo-2D HMM is the classifier.

PCA derived features have been shown to be sensitive to changes in the illumination direction [7] causing rapid degradation in verification performance. A study by Zhang et al. [8] has shown a system employing 2-D Gabor wavelet derived features to be robust to changes in the illumination direction. However, a different study by Adini et al. [9] shows that the 2-D Gabor wavelet derived features are indeed sensitive to the illumination direction.

Belhumeur et al. [7] proposed robust features based on Fisher's Linear Discriminant. However, to achieve robustness, Belhumeur's system required face images with varying illumination for training purposes.

As will be shown, 2-D DCT based features are also sensitive to changes in the illumination direction. In this paper we introduce four new techniques, which are significantly less affected by an illumination 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 2-D 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 2-D Gabor wavelets.

The rest of the paper is organized as follows. In Section 2 we briefly review the 2-D DCT feature extraction technique and describe the proposed feature extraction methods. In Section 3 we describe a Gaussian Mixture Model (GMM) classifier which shall be used as the basis for experiments. In Section 4 we describe the VidTIMIT audio-visual database. The performance of feature extraction techniques is compared in Section 5. The results are discussed and conclusions drawn in Section 6.

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

2. FEATURE EXTRACTION

2.1. 2-D Discrete Cosine Transform (DCT)

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

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

for $v,u = 0,1,2,...,N-1$

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

$$\text{and} \quad \beta(y,x,v,u) = \cos\left[\frac{(2y+1)v\pi}{2N}\right] \cos\left[\frac{(2x+1)u\pi}{2N}\right] \qquad (3)$$

The coefficients are ordered according to a zig-zag pattern, reflecting the amount of information stored [10] (see Fig. 2). For block

Appears in: Proceedings of IEEE International Conference on Image Processing, Rochester, 2002, pp. 997-1000 (Vol. 3)

located at (b, a), the DCT feature vector is composed of:

$$\left[\begin{array}{cccc} c_0^{(b,a)} & c_1^{(b,a)} & \ldots & c_{M-1}^{(b,a)} \end{array} \right]^T \quad (4)$$

where $c_n^{(b,a)}$ denotes the n-th DCT coefficient and M is the number of retained coefficients.

2.2. 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 [11].

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

$$\Delta^h c_n^{(b,a)} = \frac{\sum_{k=-K}^{K} k h_k \; c_n^{(b,a+k)}}{\sum_{k=-K}^{K} h_k k^2} \quad (5)$$

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

$$\Delta^v c_n^{(b,a)} = \frac{\sum_{k=-K}^{K} k h_k \; c_n^{(b+k,a)}}{\sum_{k=-K}^{K} h_k k^2} \quad (6)$$

where h is a $2K + 1$ dimensional symmetric window vector. In this work we shall use $K = 1$ and a rectangular window.

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 - eg. $X = I_X + I_N$. Moreover, let us also suppose that all of the blocks are corrupted with the same noise (a reasonable assumption if the blocks are small and are close or overlapping). To find the deltas for block Y, we apply Eqn. (5) to obtain (ignoring the denominator):

$$\begin{aligned} \Delta^h Y &= -X + Z & (7) \\ &= -(I_X + I_N) + (I_Z + I_N) & (8) \\ &= I_Z - I_X & (9) \end{aligned}$$

ie. 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 DCT coefficients from each block, the delta vector is $2M$ dimensional. We shall term this feature extraction method as *DCT-delta*. We interpret these delta coefficients as transitional spatial information (somewhat akin to edges).

2.3. DCT-mod, DCT-mod2 and DCT-mod-delta

By inspecting Eqns (1) and (3), it is evident that the 0th 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 Fig. 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 DCT feature vector is M dimensional):

1. Discard the first three coefficients from the baseline DCT feature vector. We shall term this *modified* feature extraction method as *DCT-mod*.
2. Discard the first three coefficients from the baseline 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, ie.:

$$\left[\begin{array}{cccccccccc} \Delta^h c_0 & \Delta^v c_0 & \Delta^h c_1 & \Delta^v c_1 & \Delta^h c_2 & \Delta^v c_2 & c_3 & c_4 & \ldots & c_{M-1} \end{array} \right]^T \quad (10)$$

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

Thus in the *DCT-mod-delta* and *DCT-mod2* approaches transitional spatial information is combined with local texture information.

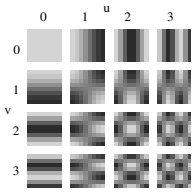

Fig. 1. Several DCT basis functions for N=8. Lighter colours represent larger values.

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

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

3. 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_G-Gaussian GMM is trained using a *k*-means clustering algorithm followed by 10 iterations of the Expectation Maximization (EM) algorithm [12].

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:

$$\mathcal{L}(X|\lambda_C) = \frac{1}{N_V} \sum_{i=1}^{N_V} \log p(\vec{x}_i|\lambda_C) \quad (11)$$

$$\text{where} \quad p(\vec{x}|\lambda) = \sum_{j=1}^{N_G} m_j \, \mathcal{N}(\vec{x}; \vec{\mu}_j, \mathbf{\Sigma}_j) \quad (12)$$

$$\text{and} \quad \lambda = \{m_j, \vec{\mu}_j, \mathbf{\Sigma}_j\}_{j=1}^{N_G} \quad (13)$$

Here λ_C is the model for person C. N_G is the number of Gaussians, m_j is the weight for Gaussian j (with constraint $\sum_{j=1}^{N_G} m_j = 1$), and $\mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma})$ is a multi-variate Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$. Given a set $\{\lambda_b\}_{b=1}^{B}$ of B background person models for person C, the average log likelihood of the claimant being an impostor is found using:

$$\mathcal{L}(X|\lambda_{\overline{C}}) = \log \left[\frac{1}{B} \sum_{b=1}^{B} \exp \mathcal{L}(X|\lambda_b) \right] \quad (14)$$

The set of background person models is found using the method described in [13]. An opinion on the claim is found using:

$$\Lambda(X) = \mathcal{L}(X|\lambda_C) - \mathcal{L}(X|\lambda_{\overline{C}}) \quad (15)$$

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. VIDTIMIT AUDIO-VISUAL DATABASE

The VidTIMIT database, created by the authors, 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 [14]. 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 mean duration of each sentence is 4.25 seconds, or approximately 106 video frames.

The recording was done in a noisy 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 384×512 pixels. The corresponding audio is stored as a mono, 16 bit, 32 kHz WAV file. For more information on the database, please visit *http://spl.me.gu.edu.au/vidtimit/* or contact the authors.

5. EXPERIMENTS

Before feature extraction can occur, the face must first be located [15]. Furthermore, to account for varying distances to the camera, a geometrical normalization must be performed. 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 of varying dimensions. Using the distance between the eyes as a size measure, an affine transformation is used [10] 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 work by Samaria [16] and Belhumeur [7]).

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

For Gabor features, we follow Duc [3] where 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.

To reduce the computational burden during modeling and testing, every second video frame was used. For each feature extraction method, 8-Gaussian client models (GMMs) were generated from features extracted from face windows in Session 1.

An artificial illumination change was introduced to face windows extracted from Sessions 2 and 3. To simulate more illumination on the left side of the face and less on the right, a new face window $v(y, x)$ is created by transforming $w(y, x)$ using:

$$v(y, x) = w(y, x) + mx + \delta \tag{16}$$

$$\text{for} \quad y = 0, 1, ..., 55 \quad \text{and} \quad x = 0, 1, ..., 63$$

$$\text{where} \quad m = \frac{-\delta}{63/2} \tag{17}$$

$$\text{and} \quad \delta = \text{illumination delta (in pixels)}$$

Example face windows for various δ are shown in Fig. 3. It must be noted that the above artificial illumination change is rather restrictive as it does not cover all the effects of illumination changes possible in real life (shadows, etc.).

Fig. 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 [13], 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 Equal Error Rate (EER) (ie. where the False Acceptance Rate is equal to the False Rejection Rate).

In the first experiment, we found the performance of the DCT approach on face windows with $\delta = 0$ (ie. no illumination change) while varying the dimensionality of the feature vectors. The results are presented in Fig. 4. 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 DCT feature vectors - hence the dimensionality of *DCT-delta* 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 DCT and all of the proposed techniques for increasing δ. Results are shown in Fig. 5.

In the third experiment we compared the performance of PCA, DCT, Gabor and *DCT-mod2* features for varying δ. Results are presented in Fig. 6.

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 (eg. β 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.

6. DISCUSSION AND CONCLUSIONS

We can see in Fig. 4 that the first three DCT coefficients contain a significant amount of person dependent information. Thus ignoring them (as in *DCT-mod*) implies a reduction in performance.

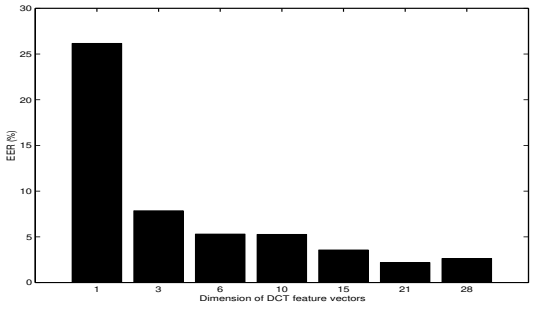

Fig. 4. Performance for varying dimensionality of DCT feature vectors

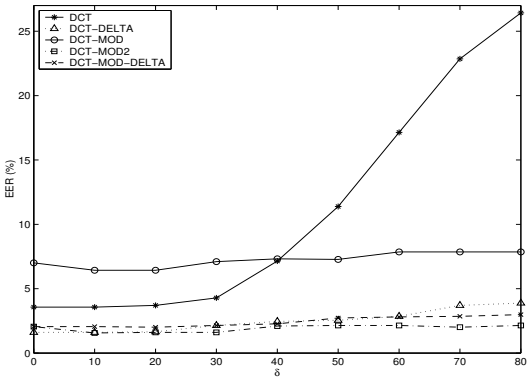

Fig. 5. Performance of DCT and proposed feature extraction techniques

This is verified in Fig. 5 where the *DCT-mod* features have worse performance than DCT features when there is little or no illumination change ($\delta \leq 30$). Performance of DCT features is fairly stable for small illumination changes but degrades for $\delta \geq 40$. This is in contrast to *DCT-mod* features which have a relatively static performance.

The remaining proposed features (*DCT-delta*, *DCT-mod-delta* and *DCT-mod2*) do not have the performance penalty present in *DCT-mod*. Moreover, all of them have similarly better performance than DCT features. *DCT-mod2* edges out *DCT-delta* and *DCT-mod-delta* in terms of stability for large illumination changes ($\delta \geq 50$). Additionally, the dimensionality of *DCT-mod2* is lower than *DCT-delta* and *DCT-mod-delta*.

The results suggest that delta features make the system more robust as well as improve performance. The results also suggest that it is only necessary to use deltas of coefficients representing the DC level and low frequency features (ie. the 0th, 1st and 2nd 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.

Comparing PCA, DCT, Gabor and *DCT-mod2* (Fig. 6), we can see that the *DCT-mod2* approach is the most immune to illumination changes - the performance is virtually flat for varying δ. The performance of PCA derived features rapidly degrades as δ increases. Performance of Gabor features is stable for $\delta \leq 40$ and then gently deteriorates as δ increases. The results suggests that we can order the features, based on their robustness and performance, as follows: *DCT-mod2*, Gabor, DCT, and lastly, PCA.

It must be noted that using the introduced illumination change, the center portion of the face (column wise) is largely unaffected.

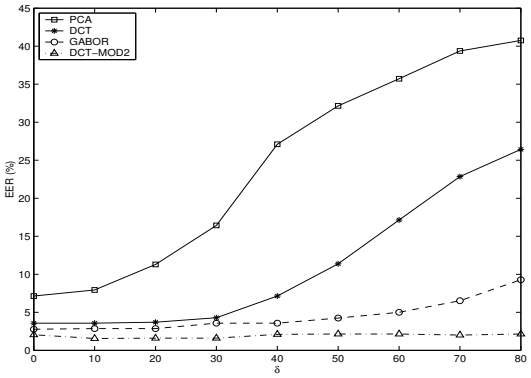

Fig. 6. Performance of PCA, DCT, Gabor and *DCT-mod2* feature extraction techniques

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, gcc 2.96)

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 (Gabor, 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 virtually unchanged, automatically leading to a degree of robustness.

It must also be noted that when using the GMM classifier in conjunction with the Gabor, DCT or *DCT-mod2* features, the spatial relation between major face features (eg. eyes and nose) is lost. However, excellent performance is still obtained.

In Table 1 we can see that 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 Gabor wavelets as well as the need to compute both real and imaginary inner products. Compared to Gabor features, PCA, DCT and *DCT-mod2* features take a relatively similar amount of time to process one face window.

7. REFERENCES

[1] G. R. Doddington et al., "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective", *Speech Communication*, Vol. 31, No. 2-3, 2000.

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

[3] B. Duc et al., "Face Authentication with Gabor Information on Deformable Graphs", *IEEE Trans. Image Proc.*, Vol. 8, No. 4, 1999.

[4] S. Lawrence et al., "Face Recognition: A Convolutional Neural-Network Approach", *IEEE Trans. Neural Net.*, Vol. 8, No. 1, 1997.

[5] S. Eickler et al., "Recognition of JPEG Compressed Face Images Based on Statistical Methods", *Image and Vision Computing*, Vol. 18, No. 4, 2000.

[6] T. S. Lee, "Image Representation Using 2D Gabor Wavelets", *IEEE Trans. Patt. Anal. and Machine Intell.*, Vol. 18, No. 10, 1996.

[7] P. N. Belhumeur et al., "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. Patt. Anal. and Machine Intell.*, Vol. 19, No. 7, 1997.

[8] J. Zhang et al., "Face Recognition: Eigenface, Elastic Matching, and Neural Nets", *Proceedings of the IEEE*, Vol. 85, No. 9, 1997.

[9] Y. Adini et al., "Face Recognition: The Problem of Compensating for Changes in Illumination Direction", *IEEE Trans. Patt. Anal. and Machine Intell.*, Vol. 19, No. 7, 1997.

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

[11] F. K. Soong and A. E. Rosenberg, "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", *IEEE Trans. ASSP*, Vol. 36, No. 6, 1988.

[12] T. K. Moon, "Expectation-maximization Algorithm", *IEEE Signal Processing Magazine*, Vol. 13, No. 6, 1996.

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

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

[15] L-F. Chen et al., "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.

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