Face recognition using holistic Fourier invariant features

Jian Huang Lai\textsuperscript{a,b}, Pong C. Yuen\textsuperscript{a,*}, Guo Can Feng\textsuperscript{a,b}

\textsuperscript{*}Department of Computer Science, Hong Kong Baptist University, Hong Kong
\textsuperscript{b}Department of Mathematics, Zhongshan University, People’s Republic of China

Received 19 March 1999; received in revised form 29 July 1999; accepted 25 August 1999

Abstract

This paper presents a new method for holistic face representation, called spectroface. Spectroface representation combines the wavelet transform and the Fourier transform. We have shown that by decomposing a face image using wavelet transform, the low-frequency face image is less sensitive to the facial expression variations. This paper also proves that the spectroface representation is invariant to translation, scale and on-the-plane rotation. To handle the rotation in depth, multiple view images are used to determine the reference image representation. Based on the spectroface representation, a face recognition system is designed and developed. Yale and Olivetti face databases are selected to evaluate the proposed system. These two databases contain 55 persons with 565 face images at different orientations, scale, facial expressions, small occlusions and different illuminations. The recognition accuracy is over 94%. If we consider the top three matches, the accuracy is over 98%. The recognition system is developed on Pentium 200 MHz computer and the recognition time is less than 3 seconds for database with 55 persons (© 2000 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Face recognition; Fourier transform; Wavelet transform; Spectroface

1. Introduction

Research on face recognition technology has been studied for more than 20 years. It has become a new major research area in the last few years because of the number of potential applications. These applications range from security access control, personal identification to human–computer communication.

A number of face recognition methods have been proposed. These methods can be divided into two approaches [1,2], namely, constituent- and face-based.

In the constituent-based approach, recognition is based on the relationship between human facial features such as eye(s), mouth, nose, profile silhouettes and face boundary [3–5]. Deformable template is a common method used in detecting facial features. Yuille et al. [3] first proposed using deformable template for eye detection in 1992. Along the Yuille et al.’s direction, Xie et al. [6] modified the eye template and the energy function for eye detection. Okada et al. [7] proposed to use Gabor wavelet to extract facial features, called JETS. Lam and Yan [5] proposed the concept of eye corners to guide the template matching process. Brunelli et al. [8] proposed integration projections to extract the facial feature which is an extension of Kanade’s work [9]. Huang et al. [10] used chrominance information to locate face and facial features. Once facial features are detected, classification can be performed using the inter-feature distance [11], Dempster–Shafer theory [12], random graph [13] or elastic graph matching technique [14]. Constituent-based approach provides high flexibility in handling non-rigid facial features, such as eyes and mouth. On the other hand, the success of this approach highly relies on the accuracy of the facial feature detection scheme. Moreover, as mentioned by Yuen et al. [15] and Feng et al.
the image resolution has to be relatively high in order to detect the facial features accurately. In turn, the computational load will be increased.

Face-based approach [17–22] attempts to capture and define the face as a whole. The face is treated as a two-dimensional pattern of intensity variation. Under this approach, the face is matched through finding its underlying statistical regularities. Principal component analysis (PCA) is a typical face-based technique. Sirovich and Kirby [17,23] first proposed to use Karhunen–Loève transform to represent human faces. In their method, faces are represented by a linear combination of weighted eigenvector. In 1991, Turk and Pentland [18] developed a face recognition system using PCA (K–L expansion). Along this direction, many PCA-based recognition system has been developed [19,24]. Theoretically, face-based (PCA-based) approach has a number of practical advantages. First, as this approach considers face as a global feature, detection of local feature is not required. Second, the resolution of the operating image can be relatively low, which will reduce the computational load. However image registration is required in PCA-based approach. This approach gives a very good result if the face to be recognized is aligned properly. Lai et al. [22] have demonstrated that the recognition accuracy using PCA falls from 85.5, 81.7, 76.7, 70.0–67.5% if the alignment is deviated from 0 to 4 pixels error. Connectionist method [25] also belongs to this approach. In 1998, Lam and Yan [20] developed an analytic-to-holistic approach for frontal-view face recognition.

To overcome the alignment problem in face-based approach, we propose to transform a face image into a feature vector, which is invariant to translation, scale and on-the-plane rotation. In this way, the limitation of face-based approach can be solved. We develop a new face representation namely, spectroface. Spectroface representation is based on the wavelet and Fourier transforms. Section 2 discusses the effect of applying wavelet transform on a face image, which indeed attenuates the impact of facial expression. We have also developed a new holistic Fourier invariant features (HFIF) which are invariant to face translation, scale and on-the-plane rotation. The detailed derivation of the HFIF is given in Section 3. A complete recognition system based on the spectroface representation is developed and reported in Section 4. Section 5 presents the experimental results of the spectroface representation and the performance of the recognition system. Conclusions are given in Section 6.

2. Extract facial expression insensitive features using wavelet transform

Earlier studies [26,27] concluded that information in low spatial frequency bands play a dominant role in face recognition. In 1986, Sergent [28] shows that the low-frequency band and high-frequency band play different roles. The low-frequency components contribute to the global description, while the high-frequency components contribute to the finer details required in the identification task. Sergent has also demonstrated that as human face is a non-rigid object, it has abundant facial expressions; and expressions influence local spatial components of face.

Nastar et al. [29,30] have investigated the relationship between variations in facial appearance and their deformation spectrum. They found that facial expressions and small occlusion affect the intensity manifold locally. Under frequency-based representation, only high-frequency spectrum is affected, called high-frequency phenomenon. Moreover, changes in pose or scale of a face affect the intensity manifold globally, in which only their low-frequency spectrum is affected, called low-frequency phenomenon. Only a change in face will affect all frequency components. Bow [31] demonstrates that the quality of the reconstructed image is very good if the image is restored with the lower half-frequency spectrum.

The above statements show that (1) the effect of different facial expressions can be attenuated by removing the high-frequency components and (2) the low-frequency components are only sufficient for recognition. As such, we proposed to use the wavelet transform for face image decomposition. The wavelet transform has been proven effective for image analysis and feature extraction. It represents a signal by localizing it in both time and frequency domains. Wavelet transform divides an image into four different frequency bands as shown in Fig. 1(b). By further decomposing the low-frequency image, we can perform the multiresolution analysis, which has been widely adopted in image processing [32–36]. Moreover, the 2D multiresolution analysis can be extended from 1D multiresolution analysis by using tensor product method. A brief review on 2D multiresolution analysis is given below.

2.1. Review of 2D multiresolution analysis

We assume that \( \{V_m\}_{m \in Z} \) of closed space \( V_m \subset L^2(R) \) is 1D multiresolution analysis, where, wavelet space is defined as \( W_m = (V_m)^\perp \), i.e., \( V_{m-1} = V_m \oplus W_m \). \( \varphi(x) \) is its scale function, \( \psi(x) \) is its wavelet function, \( \varphi(x) \) and \( \psi(x) \) satisfy the following equation:

\[
\varphi(x) = \sqrt{2} \sum_{m \in Z} h(n) \varphi(2x - n),
\]

\[
\psi(x) = \sqrt{2} \sum_{m \in Z} g(n) \varphi(2x - n),
\]

where \( g(n) = (-1)^n h(1 - n) \).

By multiresolution analysis, we can construct a pair orthogonal quadrature filters \( H \) and \( G \), where, \( H \) is
low-pass filter and $G$ is high-pass filter. For any 1D discrete signals $\{c_n\}_{n \in \mathbb{Z}}$, $\{d_n\}_{n \in \mathbb{Z}} \in \mathbb{R}^2$, such that

$$Hc_m = \sum_{n \in \mathbb{Z}} h(2m - n)c_n, \quad (2.3)$$

$$Gc_m = \sum_{n \in \mathbb{Z}} g(2m - n)c_n. \quad (2.4)$$

If 2D closed space $V_0^2 \subset L^2(R^2)$ is constructed by tensor product space, i.e., $V_0^2 = V_m \otimes V_m$, then $V_{m-1}^2 = V_m^2 \otimes (V_m \otimes V_m) \otimes (W_m \otimes V_m) \otimes (W_m \otimes W_m)$. Denote $W_m^1 = (V_m \otimes W_m), W_m^2 = (W_m \otimes V_m)$ and $W_m^3 = (W_m \otimes W_m)$.

We can verify that $\{V_m\}_{m \in \mathbb{Z}}$ constitute a 2D multi-resolution analysis, where the scale function is $\Phi(x, y) = \phi(x)\phi(y)$, the wavelet function is $\psi^1(x, y) = \phi(x)\psi(y)$, $\psi^2(x, y) = \psi(x)\phi(y)$ and $\psi^3(x, y) = \psi(x)\psi(y)$.

For any 2D discrete signals $\{x_{m,n}\}_{m,n \in \mathbb{Z}} \in \mathbb{R}^2$, let $c_{0,m,n} = x_{m,n}, m,n \in \mathbb{Z}$. In practice, $\{c_{0,m,n}\}_{m,n \in \mathbb{Z}}$ can be considered as the coefficients that face image $(x, y)$ unfolds in $V_0^2$. If we assume that $c_j = \{c_{j,m,n}\}_{m,n \in \mathbb{Z}}, \quad d_j = \{d_{j,m,n}\}_{m,n \in \mathbb{Z}}, \quad d_j^1 = \{d_{j,m,n}\}_{m,n \in \mathbb{Z}}$ and $d_j^2 = \{d_{j,m,n}\}_{m,n \in \mathbb{Z}}$, are the coefficients that face image $(x, y)$ unfolds in $V_0^2, W_1^2, W_2^2$ and $W_3^2$, respectively, Mallat iterative algorithm [32,33] can be constructed and is presented as follows:

$$c_{j,m,n} = (H \otimes H)(c_{j+1,m,n}) = \sum_{k,l} c_{j+1,k,l} h_{2m-k} h_{2n-l}, \quad (2.5)$$

$$d_{j,m,n} = (H \otimes G)(c_{j+1,m,n}) = \sum_{k,l} c_{j+1,k,l} h_{2m-k} g_{2n-l}, \quad (2.6)$$

$$d_{j,m,n}^1 = (G \otimes H)(c_{j+1,m,n}) = \sum_{k,l} c_{j+1,k,l} g_{2m-k} h_{2n-l}, \quad (2.7)$$

$$d_{j,m,n}^2 = (G \otimes G)(c_{j+1,m,n}) = \sum_{k,l} c_{j+1,k,l} g_{2m-k} g_{2n-l}. \quad (2.8)$$

### 2.2. Human face features in subbands

Fig. 1 shows the decomposition process by applying the 2D wavelet transform on a face image. The original image (shown in Fig. 1(a)) is decomposed into four subimages (shown in Fig. 1(b)). Similarly, we can obtain two levels of the wavelet decomposition as shown in Fig. 1(c) by applying wavelet transform on the low-frequency band sequentially.

In Fig. 1(b), the subband LL corresponds to the low-frequency components in both vertical and horizontal directions of the original image. Therefore, it is the low-frequency subband of the original image. The subband LH corresponds to the low-frequency component in the horizontal direction and high-frequency components in vertical direction. Therefore it holds the vertical edge details. Similar interpretation is made on the subbands HL and HH.

As the change of facial expressions mainly varies in eyes, mouth and other face muscles, from the image processing point of view, it involves mainly changes of edges. Let us take Fig. 1 as an example, the horizontal features of eyes and mouth are more clear than its vertical features, the subband LH can therefore depict major facial expression features. The subband HL, the vertical features of outline and nose are more clear than its horizontal features, depicts face pose features. The subband HH is therefore the most important for rigid object recognition because it depicts the structure feature of the object. But human faces indeed are non-rigid objects, the subband HH may be sensitive to the facial expressions. According to wavelet theory, the subband LL is the smoothed version of original image and the best approximation to the original image with lower-dimensional
space. It also contains highest-energy content within the four subbands. The subband LL features are insensitive to the facial expressions and small occlusion. Fig. 2 shows that the low-frequency images are nearly like two peas in three-level wavelet decomposition, regardless the expression of the two face images are quite different. Experimental results in Section 5.3 also show that LL subband gives the best performance among the four subbands at the same level using spectroface representation. Therefore, the LL subband is selected in this paper.

Besides, according to the properties of multiresolution analysis, we conclude that the low-frequency subband is optimal approximated image of the original image in lower-dimensional space. If applying $n$ level wavelet decomposition to the face image, the resolution of the low-frequency subband is $1/2^n$ of the resolution of the original image. Hence, wavelet transform is also effective for reducing the space dimension.

3. Holistic Fourier invariant features

How to extract the translation, scale and rotation invariant features of an image is an active research area in several disciplines such as computer vision, pattern recognition, image processing and applied optics at all times. Various translation, rotation and scale invariant methods such as the Fourier descriptor [37], moments [38] and Fourier-mellin filter [39] have been proposed. These methods provide good invariance theories but suffer from either computational complexity or accuracy problems. Moreover, Fourier descriptor method applies only on the object’s contour image.

In human face recognition, the image may be captured in different environment, the face to be recognized may have different orientations and scales comparing with the one in the reference library. In this section, we develop a method for determining the holistic
Fourier invariant features (HFIF) of face. It is proved that HFIF is

- invariant to translation;
- invariant to scale, and;
- invariant to on-the-plane rotation.

In fact, rotation of the face is three dimensional in space, i.e., rotation can be along the x-, y- and z-axis as shown in Fig. 3. Any rotation in space can be decomposed into the three fundamental rotation components in the x, y and z. HFIF can represent the facial features which are invariant for the rotation in the z-axis. For the rotation in the x- and the y-axis, we will select more view images for determining the reference face representation which will be discussed in Section 4.

3.1. Fourier transform and its properties

Fourier transform is an effective tool for signal analysis. For 2D image $f(x, y) \in L^2(\mathbb{R}^2)$, 2D Fourier transform is defined as

$$F(\mu, v) = \left[ f(x, y) \right]^\wedge = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{-2\pi i (\mu x + v y)} \, dx \, dy$$

(2.9)

which has the following properties.

- **Linear transform:** If $f(x, y), g(x, y) \in L^2(\mathbb{R}^2)$, $a, b \in \mathbb{C}$ (i.e. complex domain), then $af(x, y) + bg(x, y) \in L^2(\mathbb{R}^2)$, and

$$[af(x, y) + bg(x, y)]^\wedge = a[f(x, y)]^\wedge + b[g(x, y)]^\wedge.$$  

(2.10)

- **Invariant against spatial translation:** If $f(x, y) \in L^2(\mathbb{R}^2)$, then $f(x - a, y - b) \in L^2(\mathbb{R}^2)$ and $[f(x - a, y - b)]^\wedge = e^{-2\pi i (\mu a + v b)} F(\mu, v)$ so,

$$|[f(x - a, y - b)]^\wedge| = |e^{-2\pi i (\mu a + v b)} F(\mu, v)| = |F(\mu, v)| = |[f(x, y)]^\wedge|.$$  

(2.11)

- **Variant to scale:** If $f(x, y) \in L^2(\mathbb{R}^2)$, then $f(ax, ay) \in L^2(\mathbb{R}^2)$, and

$$[f(ax, ay)]^\wedge = \left( \frac{1}{x^2} \right) F\left( \frac{\mu}{x}, \frac{v}{x} \right)$$

(2.12)

where $x$ is a scale factor.

- **Variant to rotation:** Supposing that $x = r \cos \theta$, $y = r \sin \theta$, $\mu = \rho \cos \phi$, $v = \rho \sin \phi$, the Fourier transform in polar coordinate can be written as

$$F(\rho, \phi) = \left[ f(r, \theta) \right]^\wedge = \int_{0}^{2\pi} \int_{0}^{\infty} f(r, \theta) e^{-2\pi i \rho r \cos(\theta - \phi)} r \, dr \, d\theta.$$  

(2.13)

So, if image rotates $\phi$, then

$$[f(r, \theta + \phi)]^\wedge = F(\rho, \phi + \phi).$$  

(2.14)

3.2. Translation, scale and rotation invariant features

Section 3.1 shows that Fourier transform is translation invariant but not rotation and scale invariant. Casasent et al. [40] have demonstrated the use of Fourier transform in the orientation and scale optical correlation and employed those in optics application. Along Casasent et al.’s direction, we develop a method for computing holistic Fourier invariant features for face recognition. A complete proof is given. From Eqs. (2.12) and (2.14), we have

$$[f(ax, ay)]^\wedge = \left( \frac{1}{x^2} \right) F\left( \frac{\rho}{x}, \phi + \phi \right).$$  

(2.15)
As the factor \(1/\pi^2\) can be eliminated by normalization, the problem can be translated to consider the relationship between \(F(\rho/\alpha, \phi + \phi)\) and \(F(\rho, \phi)\).

If \(\lambda = \ln \rho\), we get \(F(\rho, \phi) = F(e^\lambda, \phi)\). Let \(F(e^\lambda, \phi) = G(\lambda, \phi)\), then \(F(\rho/\alpha, \phi + \phi) = G(\lambda - \ln \alpha, \phi + \phi)\).

It can be seen that \(G(\lambda - \ln \alpha, \phi + \phi)\) is obtained by translating \(G(\lambda, \phi)\). By applying Fourier transform on \(G(\lambda, \phi)\) and using property (2.11), we can show that this method is invariant to the translation, rotation and scale. The detailed proof is given below.

**Theorem 1.** If \(F(\rho, \phi) \in L^2([R_0, R_1] \times [0, 2\pi]), R_1 > R_0 > 0\), then

\[
F(\rho, \phi) = \sum_{n = -\infty}^{+\infty} \sum_{m = -\infty}^{+\infty} c_{nm} e^{i(2\pi n/L)(\ln \rho + m\phi)},
\]

where

\[
c_{nm} = \frac{1}{2\pi L} \int_{R_0}^{R_1} \int_{0}^{2\pi} F(\rho, \phi) e^{-i(2\pi n/L)(\ln \rho + m\phi)} \frac{1}{\rho} \, d\rho \, d\phi,
\]

\(L = \frac{1}{\ln R_1 - \ln R_0}\).

Besides, the amplitude values \(|c_{nm}|\) will not be changed with different scaling factors \(\alpha\) and rotation factors \(\phi\).

As everyone knows, Fourier transform on the interval \([0, 2\pi]\) can be described as below.

**Lemma 2.** If \(f(x) \in L^2[0, 2\pi]\), then \(f(x) = \sum_{n = -\infty}^{+\infty} c_n e^{inx}\)

where

\[
c_n = (1/2\pi) \int_{0}^{2\pi} f(x) e^{-inx} \, dx.
\]

Lemma 2 can be easily extended to the interval \([a, b]\), as follows.

**Lemma 3.** If \(f(x) \in L^2[a, b]\), then \(f(x) = \sum_{n = -\infty}^{+\infty} c_n e^{2\pi in(x/a - a)}\)

where

\[
c_n = \frac{1}{b - a} \int_{a}^{b} f(x) e^{-2\pi in(x/a - a)} \, dx.
\]

Lemma 3 can be verified if let \(x = [(b - a)/2\pi] y + a\) and substituting it into Lemma 2. Similarly, we can show the Fourier transform is valid on the interval \([a, b] \times [0, 2\pi]\) and we get Lemma 4.

**Lemma 4.** If \(f(\sigma, \phi) \in L^2([a, b] \times [0, 2\pi])\), then

\[
f(x, y) = \sum_{n = -\infty}^{+\infty} \sum_{m = -\infty}^{+\infty} c_{nm} e^{i(2\pi n/L)(b - a) + m\phi},
\]

where

\[
c_{nm} = \frac{1}{2\pi L} \int_{0}^{2\pi} \int_{a}^{b} f(x, y) e^{-(i(2\pi n/L)(b - a) + m\phi)} \, dx \, dy.
\]

Now, Theorem 1 is proved as follows.

**Proof of the Theorem.** Let \(\lambda = \ln \rho\) and \(L = 1/(\ln R_1 - \ln R_0)\), then \(G(\lambda, \phi) = F(e^\lambda, \phi) \in L^2([\ln R_0, \ln R_1] \times [0, 2\pi]\).

By Lemma 4, we have

\[
G(\lambda, \phi) = \sum_{n = -\infty}^{+\infty} \sum_{m = -\infty}^{+\infty} c_{nm} e^{i(2\pi n/L)(\lambda - m\phi)},
\]

where

\[
c_{nm} = \frac{1}{2\pi L} \int_{0}^{2\pi} \int_{\ln R_0}^{\ln R_1} G(\lambda, \phi) e^{i(2\pi n/L)(\lambda - m\phi)} \, d\lambda \, d\phi.
\]

Substituting \(\lambda = \ln \rho\) into Eqs. (2.16) and (2.17), we get

\[
F(\rho, \phi) = \sum_{n = -\infty}^{+\infty} \sum_{m = -\infty}^{+\infty} c_{nm} e^{i(2\pi n/L)(\ln \rho + m\phi)},
\]

where

\[
c_{nm} = \frac{1}{2\pi L} \int_{0}^{2\pi} \int_{\ln R_0}^{\ln R_1} F(\rho, \phi) e^{i(2\pi n/L)(\ln \rho + m\phi)} \frac{1}{\rho} \, d\rho \, d\phi.
\]

If the face image changes with the scaling factor \(\alpha\) and the rotation factor \(\phi\) in spatial domain, it can be represented as \(F(\rho/\alpha, \phi + \phi)\) in frequency domain. The coefficients that Fourier transform is applied to \(F(\rho/\alpha, \phi + \phi)\) are

\[
c'_{nm} = \frac{1}{2\pi L} \int_{\ln R_0}^{\ln R_1} \int_{\ln R_0}^{\ln R_1} F(\rho, \phi) e^{i(2\pi n/L)(\ln \rho + m\phi)} \frac{1}{\rho} \, d\rho \, d\phi.
\]

To show that the HFIF is orientation and scale invariant, we need to prove that the amplitude values \(|c'_{nm}|\) will not be changed with different scaling factors \(\alpha\) and rotation factors \(\phi\), i.e. \(|c'_{nm}| = |c_{nm}|\) for all \(m\) and \(n\).

Let \(\rho' = \rho/\alpha, \phi' = \phi + \phi\), then \(\rho = \alpha \rho', \phi = \phi' - \phi\), replacing \(\rho\) and \(\phi\) into Eq. (2.18), we have

\[
c'_{nm} = \left|c_{nm}\right| = \frac{1}{2\pi L} \int_{\ln R_0}^{\ln R_1} \int_{\ln R_0}^{\ln R_1} F(\rho', \phi') e^{i(2\pi n/L)(\ln \rho' + m\phi')} \frac{1}{\rho'} \, d\rho' \, d\phi'.
\]

Hence, \(|c'_{nm}| = |c_{nm}|\).
4. Face recognition system based on spectroface representation

A standard face recognition system includes the following steps:
- extract human face from image;
- represent a human face;
- classification of face.

In the proposed system, it is assumed that the human face has been extracted from an image. Various matured techniques that locate the face from image have been published [41,42]. If the input is a colour image, detection based on skin colour [43] can be used.

4.1. What is spectroface?

Spectroface representation is based on the wavelet transform and holistic Fourier invariant features. Wavelet transform is applied to the face image in order to eliminate the effect of different facial expressions. Also, decomposing the face image will reduce the resolution of the image, which in turn, reduce the computation of the recognition system. Details for the wavelet transform have been discussed in Section 2. After decomposing the face image, we extract the holistic Fourier invariant features (HFIF) from the low-frequency subband image. Section 3 has proved that the HFIF is translation, scale and on-the-plane rotation invariant. The block diagram of the spectroface representation is shown in Fig. 4.

The spectroface representation algorithm is described as follows:

1. Decompose the face by the wavelet transform until the resolution of the low-frequency subband is 32 × 32. This resolution is chosen because Lai et al. [22] have performed an experiment on the effect of the low-frequency wavelet subband against the recognition accuracy. For most of the cases, resolution 32 × 32 and 16 × 16 give the best results while in some cases, 16 × 16 resolution gives lower accuracy than 32 × 32 resolution. Therefore, the resolution 32 × 32 is used. The face image is then represented by the low-frequency subband.

2. Apply FFT to the low-frequency subband. The FFT coefficients are invariant to the spatial translation.

3. Using the center of the FFT map (as shown in Fig. 4), the FFT map is represented by polar coordinate.

Making a mapping of $\lambda = \ln \rho$, the second FFT is applied to the resultant image in step (3). The output is the spectroface representation, which is invariant to translation, scale and on-the-plane rotation.

4.2. Proposed face recognition system

The proposed face recognition system using spectroface consists of two stages, namely, training stage and recognition stage. Training stage represents a set of reference images as spectrofaces, and stores them into a gallery (reference library). Recognition stage translates a probe image into spectroface representation, and then matches it with those reference images stored in the gallery to identify the face image. Fig. 5 shows the system block diagram. The spectroface representation has been discussed in Section 4.1. The other building blocks are discussed as follows.

4.2.1. Preprocessing

Preprocessing includes two steps, namely (1) extract face from image and (2) illumination enhancement. As
many face detection algorithms have been developed, this system assumes that the face region has been segmented. This section mainly highlights how the illumination can be enhanced.

As mentioned in most of the literatures, different illumination conditions affect the performance of most of the face recognition system. Basically, it can be divided into four illumination conditions; (1) the normal illumination condition; (2) uniform illumination with linear scale of lighting intensity; (3) uniform illumination with non-linear scale of lighting intensity; (4) non-uniform illumination (light source may be from different directions). Examples of four different cases are shown in Fig. 6.

The second case can be adjusted by normalizing the whole image intensity. If \( f(n, m) \), where \( 0 \leq n \leq N \) and \( 0 \leq m \leq M \), represents the face image, the normalized face image is then represented by \( I(n, m) \) which is defined as

\[
I(n, m) = \frac{f(n, m)}{\|f\|}, \quad 0 \leq n \leq N, \quad 0 \leq m \leq M, \tag{4.1}
\]

where

\[
\|f\| = \sqrt{\sum_{n=0}^{N} \sum_{m=0}^{M} f(n, m)^2}.
\]

In the third case, if the probe image is in uniform illumination with non-linear scale of lighting intensity, the gray-level intensity of the image can be adjusted using gray-level histogram equalization [31].

The most difficult situation is the last case in which the light source may come from different directions. Basically, it is hard to find a generic model to solve this case. Instead, we propose to represent the reference face image feature vector, which will be stored in the gallery, by images with different illuminations. Details will be discussed in Section 4.2.2. In summary, the preprocessing algorithm is given below:

1. apply gray-level histogram equalization to the face image;
2. normalize the intensity of the face image using Eq. (4.1).

4.2.2. Reference image representation

It has been proved that spectroface representation is invariant to the translation, scale and on-the-plane rotation. However, in practice, the face to be recognized may have a rotation in depth (i.e. rotation around the \( x \)- and the \( y \)-axis in Fig. 3). Whenever there is a rotation in depth, the face information actually has been lost. This information cannot be recovered based on a single image. To handle the variations due to rotation in depth, two
methodologies have been proposed. The first approach treats every view as a single problem, called view-based approach. The results of this approach are good if the orientation of the probe face image can be detected. The second approach encodes both identity as well as viewing conditions, called parametric approach. The advantages of this approach are (1) its simplicity and (2) that there is no need to estimate the orientation of the probe face in advance. However, since each viewing condition of the same identity is considered, the complexity of this approach will be \( N \) times greater than that of the view-based approach, where \( N \) is the number of views to be considered for each identity.

In the proposed system, HFIF is based on Fourier transform, which has a linear property. That means, more than one face images can be weighted as a reference face representation. In this way, even though we consider more view images to construct the reference face representation, only one representation per identity needs to be stored in the gallery. Such representation has already considered the face image with different viewing conditions. In this way, the complexity would not be increased. Using the same argument, images with different illumination conditions are also considered.

The representation for each person \( V_i \) is calculated from different viewing and illumination conditions of the same person and is given by

\[
V_i = \frac{1}{N} \sum_{j=1}^{N} S_{ij},
\]

(4.2)

where \( S_{ij} \) is the spectroface representation of the \( j \)th view image of the \( i \)th person.

4.2.3. Similarity measurement

A typical and popular Euclidean distance is employed to measure the similarity between two vectors.

4.3. The computational complexity of spectroface

Suppose there are \( N \) reference images. The resolution of each image is \( d \) \((d \) is the number of pixels in an image). The complexity of the wavelet transform is then \( O(d \log d) \). Suppose the resolution of the low-frequency subband is \( d'(d' < d) \), the complexity of applying Fourier transform on the low-frequency subband is \( O(d' \log d') \). Finally, the complexity for matching \( N \) reference images is \( O(Nd') \). Therefore, the computational complexity for spectroface is \( O(Nd' \log d + 2Nd' \log d') \) in the training stage and \( O(d' \log d + 2d \log d + Nd) \) in the recognition stage.

For the eigenface method, the computation complexity of PCA is \( O(d^3 + Nd^2) \) in the training stage and \( O(d^2 + Nd) \) in the recognition stage.

Take the training of the Olivetti database on a Pentium 200 MHz PC as an example, the Eigenface method takes 23.3 min while Spectroface takes only 8.8 min.

5. Experiment results and discussions

The experimental results presented in this section are divided into five parts. The first part presents the implementation details of the proposed method. The second part evaluates the translation, scale and on-the-plane rotation invariant properties of the spectroface representation. Part three compares the performance of different wavelet subbands using spectroface representation. The fourth part focuses on using images with multiple viewing and illumination conditions to construct the reference image representation. Finally, a comparison between the proposed method and the existing methods are presented.

5.1. Implementation details

5.1.1. Face image databases

Two standard databases from Yale University and Olivetti research laboratory are selected to evaluate the recognition accuracy of the proposed system. These databases include face images with different expressions, small occlusion, different illumination condition and different poses.

In the Yale database, there are 15 persons and each person consists of 11 different facial views that represent various expressions, illumination conditions and small occlusion (by glasses). One of the persons is shown in Fig. 7. Hence, there are 165 face images in the database. The resolution of all images is 128 \( \times \) 128 (Fig. 8).

In the Olivetti database, there are 40 persons and each person consists of 10 different facial views that represent various expressions, small occlusion (by glasses), different scale and orientations. Hence, there are 400 face images in the database. One of the persons is shown in Fig. 9. The resolution of all images is 128 \( \times \) 128.

5.1.2. Wavelet function

Generally speaking, all wavelet transforms with smooth, compactly support, orthogonality (or biorthogonality) can be used in Spectroface method. It is found that the effect of different wavelet does not affect the performance of spectroface. Throughout this paper, the well-known Daubechies wavelet D4 [34,35] is adopted and its filter coefficients are:

\[
\begin{align*}
 h_0 & = 4.8296291314453416E - 01, \\
 g_0 & = -1.2940952255126037E - 01, \\
 h_1 & = 8.3651630373780794E - 01, \\
 g_1 & = -2.241386804201339E - 01, \\
 h_2 & = 2.241386804201339E - 01, \\
 g_2 & = 8.3651630373780794E - 01, \\
 h_3 & = -1.2940952255126037E - 01, \\
 g_3 & = -4.8296291314453416E - 01.
\end{align*}
\]
Fig. 7. Various view images of one person in the Yale database.

Fig. 8. nor.image in Fig. 7 is rotated with different degrees.

Fig. 9. Various view images of one person in Olivetti database.
5.1.3. Definition of terms

5.1.3.1. Recognition accuracy. The output of a recognition system is a list of sorted reference images in descending order by similarity with the testing (probe) image. That means, the reference image on the top of the list has the highest similarity (lowest distance) with the testing image. Recognition accuracy is defined as follows:

\[
\text{Recognition accuracy} = \frac{\text{Number of corrected matches}}{\text{Total Number of testing images}} \times 100\%.
\]

The corrected matching means that the reference image on the top of the list is the testing image.

5.1.3.2. Rank \(k\) recognition accuracy. Another parameter is used to measure the system performance, called rank \(k\) recognition accuracy or rank(\(k\)). In rank(\(k\)), matching is corrected if one of the top \(k\) reference images on the list is the testing image. Therefore, rank(1) is equal to the recognition accuracy defined in Section 5.1.3.1, and rank(\(k\)) \(\geq\) rank(\(n\)) if \(k \geq n\). This figure actually reflects the cumulative accuracy.

5.2. Evaluate the invariant properties of spectroface

This section demonstrates the invariant properties of the spectroface representation. A test database is constructed as follows. In Yale database, each person has 12 different facial expressions. Each image is then rotated with from \(-30^\circ\) to \(+30^\circ\) and \(+90^\circ\). The rotated images of one person are shown in Fig. 8.

Following the algorithm described in Section 4.1, applying spectroface method on the image shown in Fig. 10(a), (b) and (c), the partial spectroface coefficients are shown in Tables 1–3, respectively. The bold and italic entry in each table represents the center of the spectroface representation. The entry away from the center will have a smaller value. The partial datum in each table contributes more than 95% of the total energy of the vector so these can be used as representative. It can be seen that even though the image is rotated, translated and scaled, there is only a small variation on the coefficients. In fact, the Euclidean distances between (a) and (b), and (a) and (c) are 5.0929E-005 and 1.7554E-004, respectively (other data out of the table are also included). These results

---

**Table 1**
The partial datum that around the center of feature matrix of nor.image

<table>
<thead>
<tr>
<th></th>
<th>0.0055</th>
<th>0.0051</th>
<th>0.0055</th>
<th>0.0063</th>
<th>0.0051</th>
<th>0.0052</th>
<th>0.0054</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0630</td>
<td>0.0644</td>
<td>0.0766</td>
<td>0.0755</td>
<td>0.0807</td>
<td>0.0754</td>
<td>0.0474</td>
<td></td>
</tr>
<tr>
<td>0.0055</td>
<td>0.0052</td>
<td>0.0051</td>
<td>0.0064</td>
<td>0.0056</td>
<td>0.0051</td>
<td>0.0055</td>
<td></td>
</tr>
<tr>
<td>0.1479</td>
<td>0.1714</td>
<td>0.2733</td>
<td>0.4931</td>
<td>0.2733</td>
<td>0.1714</td>
<td>0.1479</td>
<td></td>
</tr>
<tr>
<td>0.0055</td>
<td>0.005</td>
<td>0.0055</td>
<td>0.0064</td>
<td>0.0051</td>
<td>0.0052</td>
<td>0.0055</td>
<td></td>
</tr>
<tr>
<td>0.0474</td>
<td>0.0754</td>
<td>0.0807</td>
<td>0.0755</td>
<td>0.0766</td>
<td>0.0644</td>
<td>0.063</td>
<td></td>
</tr>
<tr>
<td>0.0054</td>
<td>0.0052</td>
<td>0.0051</td>
<td>0.0063</td>
<td>0.0055</td>
<td>0.0051</td>
<td>0.0055</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2**
The partial datum that around the center of feature matrix of nor.image rotated 90°

<table>
<thead>
<tr>
<th></th>
<th>0.0055</th>
<th>0.0052</th>
<th>0.0051</th>
<th>0.0064</th>
<th>0.0051</th>
<th>0.0052</th>
<th>0.0054</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0631</td>
<td>0.0642</td>
<td>0.0763</td>
<td>0.0772</td>
<td>0.0806</td>
<td>0.0751</td>
<td>0.0471</td>
<td></td>
</tr>
<tr>
<td>0.0054</td>
<td>0.0051</td>
<td>0.0055</td>
<td>0.0064</td>
<td>0.0051</td>
<td>0.0051</td>
<td>0.0055</td>
<td></td>
</tr>
<tr>
<td>0.1483</td>
<td>0.1719</td>
<td>0.2739</td>
<td>0.4915</td>
<td>0.2739</td>
<td>0.1719</td>
<td>0.1483</td>
<td></td>
</tr>
<tr>
<td>0.0055</td>
<td>0.0051</td>
<td>0.0055</td>
<td>0.0064</td>
<td>0.0055</td>
<td>0.0051</td>
<td>0.0054</td>
<td></td>
</tr>
<tr>
<td>0.0471</td>
<td>0.0751</td>
<td>0.0806</td>
<td>0.0772</td>
<td>0.0763</td>
<td>0.0642</td>
<td>0.0631</td>
<td></td>
</tr>
<tr>
<td>0.0054</td>
<td>0.0051</td>
<td>0.0056</td>
<td>0.0064</td>
<td>0.0051</td>
<td>0.0052</td>
<td>0.0055</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3**
The partial datum that around the center of feature matrix of nor.image rotated +10°

<table>
<thead>
<tr>
<th></th>
<th>0.0048</th>
<th>0.0043</th>
<th>0.0055</th>
<th>0.0062</th>
<th>0.0041</th>
<th>0.0041</th>
<th>0.0052</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0645</td>
<td>0.0574</td>
<td>0.0605</td>
<td>0.0746</td>
<td>0.0744</td>
<td>0.0703</td>
<td>0.0494</td>
<td></td>
</tr>
<tr>
<td>0.0051</td>
<td>0.0040</td>
<td>0.0040</td>
<td>0.0061</td>
<td>0.0054</td>
<td>0.0042</td>
<td>0.0048</td>
<td></td>
</tr>
<tr>
<td>0.1639</td>
<td>0.2010</td>
<td>0.2977</td>
<td>0.4994</td>
<td>0.2977</td>
<td>0.2010</td>
<td>0.1639</td>
<td></td>
</tr>
<tr>
<td>0.0048</td>
<td>0.0042</td>
<td>0.0054</td>
<td>0.0061</td>
<td>0.0040</td>
<td>0.0040</td>
<td>0.0051</td>
<td></td>
</tr>
<tr>
<td>0.0494</td>
<td>0.0703</td>
<td>0.0744</td>
<td>0.0746</td>
<td>0.0605</td>
<td>0.0574</td>
<td>0.0645</td>
<td></td>
</tr>
<tr>
<td>0.0052</td>
<td>0.0041</td>
<td>0.0041</td>
<td>0.0062</td>
<td>0.0055</td>
<td>0.0043</td>
<td>0.0048</td>
<td></td>
</tr>
</tbody>
</table>
Table 4
Performance comparison using Yale database on four subbands

<table>
<thead>
<tr>
<th>Subband</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>95.00%</td>
<td>41.67%</td>
<td>45.00%</td>
<td>26.67%</td>
</tr>
</tbody>
</table>

Test images do not include lef.image and rig.image
Test images include both lef.image and rig.image

Table 5
Performance comparison using Olivetti database on four subbands

<table>
<thead>
<tr>
<th>Subband</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition accuracy</td>
<td>81.94%</td>
<td>17.22%</td>
<td>29.72%</td>
<td>2.78%</td>
</tr>
</tbody>
</table>

Table 6
Recognition accuracy when reference face image representation is formed by 3 view images

<table>
<thead>
<tr>
<th>Rank(1)</th>
<th>Rank(2)</th>
<th>Rank(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>94.64%</td>
<td>98.21%</td>
<td>99.28%</td>
</tr>
</tbody>
</table>

Fig. 11. Two-level wavelet decomposition.

5.3. Performance comparison of different wavelet subbands

This section evaluates the recognition performance of four subbands at the same wavelet decomposition level using Spectroface representation. For Yale database, nor.image in Fig. 7 is selected as reference image in gallery (total 15 images) and the rest are selected as probe image set (total 150 probe images). For Olivetti database, we select the first image in Fig. 9 (for each person) as reference image in gallery (total 40 images) and the rest of the images form a probe image set (total 360 probe images).

A two-level wavelet decomposition is applied on each image (shown in Fig. 11). Subbands 1, 2, 3 and 4 in Fig. 11 are selected for evaluation and the recognition accuracy for each subband is calculated and tabulated in Tables 4 and 5. For both databases, subband 1 (i.e. subband LL in Fig. 1) gives the best result. This result justifies the use of subband LL in our spectroface representation.

Moreover, the recognition accuracy for the proposed method is 91.33% for Yale database, while it is 95.0% if the two poor illumination images are eliminated. The recognition accuracy is 81.94% for Olivetti database. These results together show that the proposed method is excellent for the frontal-view images with different facial expressions, on-the-plane rotations and small occlusion. The proposed method is not for handling face with rotation in-depth. To handle the rotation in-depth, we will use multi-view images to represent the reference face image, which will be discussed in the next section.

5.4. Using multiple view images to represent reference face image

According to the linear property of the Fourier transform, more than one face images can be weighted as a reference face representation in the training stage. In this way, we propose to use multiple view images to form reference face representation. Basically, there are two types of face images that need to be involved in determining the reference face representation, namely face image with different poses and with different illumination. Below are two experiments using multiple view images to represent reference face image.

5.4.1. Reference face image representation formed by 3 view images

As Yale database does not contain image with different poses, only Olivetti database is used for this experiment. Images of one of the persons in Olivetti are shown in Fig. 9. We select the first three face images, i.e. normal face, slight left profile and slight right profile to form the reference face representation (total 120 training images) based on Eq. (4.2) The rest (total 320 images) are used for probe images, the recognition results are shown in Table 6. It can be seen that the accuracy is increased from 81.94 to 94.64%. If we consider the top three ranks, the accuracy is 99.28%. The results are very encouraging.
Table 7

<table>
<thead>
<tr>
<th></th>
<th>Rank(1)</th>
<th>Rank(2)</th>
<th>Rank(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test images do not include lef.image and rig.image</td>
<td>99.05%</td>
<td>99.05%</td>
<td>99.05%</td>
</tr>
<tr>
<td>Test images include lef.image and rig.image</td>
<td>95.56%</td>
<td>97.78%</td>
<td>98.52%</td>
</tr>
</tbody>
</table>

5.4.2. Reference face image representation formed by more than one illumination condition images

As the illuminations of all images are the same in Olivetti database, only Yale database is used in this experiment. Fig. 6 shows that there are four kinds of illuminations in Yale database, i.e., the images nor.image, cen.image, lef.image and rig.image. We select the 2 images, nor.image and cen.image, to construct the reference image representation (total 30 images). The rest (total 135 probe images) are used for probe images. The recognition results are shown in Table 7. It can be seen that the recognition rate increase from 91.33% (Table 4) to 95.56%. If we consider the top three matches, the accuracy is 98.52%. Moreover, if the two non-uniform lighting images are removed from the probe image set, the recognition accuracy is 99.05%.

Table 8

<table>
<thead>
<tr>
<th>Method</th>
<th>Recognition accuracy</th>
<th>Computational load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method</td>
<td>95.56%</td>
<td>Low ($O(d \cdot \log d)$)</td>
</tr>
<tr>
<td>Eigenface (from [44])</td>
<td>80.40%</td>
<td>High</td>
</tr>
<tr>
<td>Eigenface w/o 1st 3 ([44])</td>
<td>89.20%</td>
<td>High</td>
</tr>
<tr>
<td>Correlation (from [44])</td>
<td>80.00%</td>
<td>High</td>
</tr>
<tr>
<td>Linear subspace (from [44])</td>
<td>84.40%</td>
<td>High</td>
</tr>
<tr>
<td>Fisherface (from [44])</td>
<td>94.00%</td>
<td>High</td>
</tr>
</tbody>
</table>

5.5. Comparisons with other existing systems

This section compares the proposed system performance with five existing systems, namely, Eigenface [44], Fisherface [44], Correlation [44], Linear Subspace [44] and Lam et al.'s method [20] using Yale and Olivetti databases.

Lam et al. [20] have proposed a recognition system based on the analytic-to-holistic approach. According to the article, using Olivetti database, the accuracy for rank(1), rank(2) and rank(3) (the rank definition is the same as in this paper) are around 84, 90 and 96% as shown in Table 8. Comparing with the performance with the proposed method, i.e. 94.64, 98.21 and 99.21%, our method gives better performance. Moreover, in terms of computational load, Lam et al.'s method is computational expensive because detection of facial landmarks are required before classification is performed. For the proposed method, the computational load is $O(d \cdot \log d)$.

Using Yale database, Belhumeur et al. [44] have evaluated the recognition performance with four methods namely, Eigenface, correlation, Linear subspace and Fisherface [44]. The results are shown in Table 9. From [44], Fisherface gives the best performance with recognition accuracy 94%. Using the proposed method, the recognition accuracy on Yale database is 95.56%. Moreover, according to [44], all these four methods are computational expensive.

6. Conclusions

The paper presents a spectroface representation method that combines the Fourier transform and the wavelet transform for face recognition. A recognition system based on spectroface representation is designed and developed. Two standard databases from Yale University and Olivetti research laboratory are selected to evaluate the proposed system. The databases contain face image with different orientation, scale, facial expression, small occlusions and different types of illuminations. The recognition accuracy of the proposed system is over 94%. If we consider the top three matches, the accuracy is over 98%. The recognition time is less than 3 seconds for database with 55 persons on Pentium 200 MHz computer.
Our proposed system has the following additional advantages:

- Compared with the constituent-based approach, the proposed method does not require detecting any facial features, such as eyes and mouth.
- Compared with the face-based approach, the proposed spectroface representation does not require detecting any facial landmarks for alignment, as the proposed method is translation and scale invariant. It is well known that it is hard to detect facial landmarks accurately.
- Even though the proposed system uses more than one images to construct the reference image representation, which is stored in the gallery (reference library), it would not increase our recognition complexity. For each person, we store only one reference image representation in the gallery, though more than one images are used to determine such a representation.
- The proposed Holistic Fourier Invariant Features (HFIF) can be directly applied to all 2D planar objects. For such an object, HFIF are invariant for its orientation and scale.

The proposed method works very well on images with different facial expressions, translation, scale, occlusion, on-the-plane rotation, small rotation in depth ($\pm 30^\circ$) and uniform illumination conditions. The future improvement of this project will continue on developing other possibility to handle (1) large rotation in depth and (2) non-uniform illumination conditions.

Acknowledgements

This project was supported by the Faculty Research Grant, Hong Kong Baptist University, and the first author is partly supported by Natural Science Foundation of Guangdong (No: 960033). The authors would like to thank (1) Prof. D.G. Deng and Prof. D.Q. Dai of Zhongshan University for helpful discussions, (2) Yale University and Olivetti Research laboratory for the databases and (3) Ms. Joan Yuen for proofreading this manuscript.

References


About the Author—J.H. LAI received his M.Sc. degree in applied mathematics in 1989 and his Ph.D. degree in mathematics in 1999 from Zhongshan (Sun Yat-sen) University, China. Now he is an Associate Professor in the Department of Mathematics, Zhongshan (Sun Yat-sen) University, China. His current interests include human face recognition, pattern recognition and computer vision.

About the Author—P.C. YUEN was born in Hong Kong in 1966. He received his B.Sc. degree in electronic engineering with first class honours in 1989 from City Polytechnic of Hong Kong, and his Ph.D. degree in electrical and electronic engineering in 1993 from The University of Hong Kong. He is now an Associate Professor in the Department of Computer Science, Hong Kong Baptist University. His major research interests include computer vision, image analysis, human face recognition and pattern recognition.

About the Author—G.C. FENG received his M.Sc. degree in bio-mathematics from Zhongshan (Sun Yat-sen) University, China in 1988. Now he is pursuing his Ph.D degree at the Hong Kong Baptist University. His current interests include pattern recognition and computer vision.