Pattern Recognition 44 (2011) 2588–2597

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Human face recognition based on multidimensional PCA and extreme learning machine

A.A. Mohammed *, R. Minhas, Q.M. Jonathan Wu, M.A. Sid-Ahmed

Department of Electrical and Computer Engineering, University of Windsor, 401 Sunset Ave., Windsor, Ont., Canada N9B 3P4

ARTICLE INFO

Article history:
Received 7 June 2010
Received in revised form
30 November 2010
Accepted 12 March 2011
Available online 21 March 2011

Keywords:
Face recognition
Multiresolution analysis
Bidirectional two dimensional principal component analysis
Extreme learning machine
KNN classifier

ABSTRACT

In this work, a new human face recognition algorithm based on bidirectional two dimensional principal component analysis (B2DPCA) and extreme learning machine (ELM) is introduced. The proposed method is based on curvelet image decomposition of human faces and a subband that exhibits a maximum standard deviation is dimensionally reduced using an improved dimensionality reduction technique. Discriminative feature sets are generated using B2DPCA to ascertain classification accuracy. Other notable contributions of the proposed work include significant improvements in classification rate, up to hundred folds reduction in training time and minimal dependence on the number of prototypes. Extensive experiments are performed using challenging databases and results are compared against state of the art techniques.

1. Introduction

Face recognition has attracted research community during the last few decades as it is the most common visual pattern in our environment. Significant development in this area has facilitated emergence of wide range of face recognition systems for commercial and law enforcement applications. Typical applications include driver's license, passports, voter registration card, human–computer interaction, database security, law enforcement, and virtual reality. Face recognition is non-intrusive, i.e., images can be captured, identified or verified even without the knowledge and physical interaction of the subject. Moreover, an expert is not required to analyze or interpret the results and data can be easily collected using simple image acquisition devices. Humans recognize faces with natural ease, however, automated face recognition is very challenging since faces belong to a class of natural objects that do not lend themselves to simple geometric interpretations. The advantage of computer-aided face recognition is its ability to handle large number of faces; whereas, a human brain has limited memory. Despite massive intricacies, the human visual system efficiently discriminates and recognizes faces. Aging, changes in facial hair, illumination, viewpoint variations, and cluttered background are the major challenges tackled by an automatic face recognition system.

A fully automated face recognition system must reliably perform three subtasks: face detection, feature extraction and recognition/identification. In the past, the problem of automatic face recognition has been addressed in different fashions; some researchers introduced the use of localized faces for feature extraction and classification [1]. On the other hand, a few schemes isolate these subtasks to simplify automated face recognition, enhance the assessment and advancement of individual component techniques. The use of localized face portion improves classification accuracy, however, such localization requires an additional module that increases the computational complexity. Once faces are localized, the recognition task is greatly simplified since background clutter and erroneous information is eliminated. To imitate real-life scenarios, some databases are deliberately generated at different time instances in presence of cluttered background and at varying levels of scale. In such situations, manual face localization will obliterate the objective of these datasets. To rigorously test the performance of our algorithm, challenging face databases have been used without prior face localization and our principal focus is on the development of a new and efficient feature extraction method based on global image content, i.e., face and non-face portions.

Automatic face recognition systems are classified into two broad categories, namely, constituent and face based recognition [2]. Constituent face recognition is based on relationship between human facial features such as eyes, nose, mouth and

* Corresponding author.
E-mail addresses: a.mohammea@gmail.com (A.A. Mohammed), minhasr@uwindsor.ca (R. Minhas), jwu@uwindsor.ca (Q.M. Jonathan Wu), ahmed@uwindsor.ca (M.A. Sid-Ahmed).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.03.013

Fig. 1. Sample images of a subject from FERET database.

Fig. 2. Sample images of a subject from Faces94 database.

facial boundary [3]. This approach significantly relies on the accuracy of facial feature detection. Reliable extraction of facial features is an extremely complicated task since human faces have similar features with subtle changes in size and geometry that make them different from one another. Due to aforementioned complications researchers have proposed face based recognition systems [4] wherein a human face is treated as a two-dimensional intensity pattern and recognition is achieved through detection and matching of its statistical properties. In this work, we focus on face based recognition and briefly discuss some of the well established techniques in literature.

To improve speed and accuracy of a face recognition system various dimensionality reduction techniques have been developed. Kirby et al. [5] represented human faces as a linear combination of weighted eigenvectors using principal component analysis (PCA). PCA based face recognition systems suffer from poor discriminatory power and high computational load, therefore, to eliminate these limitations Bartlett et al. [6] proposed the use of independent component analysis (ICA). In [7], authors proposed the use of linear discriminant analysis (LDA) to maximize the ratio of between-class scatter matrix and the within-class scatter matrix for improved face recognition. An eigenspace based adaptive approach that utilizes a specific kind of genetic algorithm called evolutionary pursuit (EP) [8] and elastic bunch graph matching (EBGM) [9] have also been proposed to generate an optimal set of projection axes. Bach et al. [10] proposed the use of kernel Hilbert space for ICA to adaptively generate nonlinear functions and to obtain a robust algorithm with regards to variations in source density, degree of non-Gaussianity, and presence of outliers. A kernel machine-based discriminant analysis method [11] that deals with nonlinear distribution of the face pattern has also been used for improved representation of faces under variations in viewpoint, illumination and facial expression. To enhance the discriminative power of extracted features and to achieve superior face recognition researchers have also proposed the use of Bayesian framework [12,13] and support vector machines (SVM) [14,15].

Transform based approaches have been proposed to improve the performance of a face recognition system for images with high dimensionality. Face images are transformed into a new domain followed by application of PCA or other dimensionality reduction techniques. Development of enhanced multiresolution analysis techniques have encouraged research community to apply these tools to achieve a high level of class separability in pattern recognition applications. Common wavelet based face recognition architectures include wavelet based PCA [16], wavelet based LDA [17], wavelet based kernel association memory (kAM) [18] and wavelet based modular weighted PCA [19]. Emergence of curvelets [20] that offer enhanced directional and edge representation has prompted researchers to apply them to several areas of image processing. Curvelet based PCA [21], curvelet based LDA [21] and curvelet based PCA+LDA [21] are some recent curvelet based face recognition approaches. The ingrained limitations of existing face recognition algorithms include large sensitivity to viewpoint variations, number of prototypes, and slow classification speed. This work combines curvelet transform with bidirectional two dimensional principal component analysis (B2DPCA) and extreme learning machine (ELM) to eliminate the inherent shortcomings of previous methods. We performed extensive experiments to demonstrate superiority of our proposed technique over existing state-of-the-art methods. A few sample images of one of the subject from FERET [22] and Faces94 [23] database is shown in Figs. 1 and 2, respectively.

The remainder of the paper is divided into five sections. Section 2 discusses feature extraction using curvelet transform, followed by a discussion of B2DPCA in Section 3. ELM classifier is presented in Section 4 and the proposed method is detailed in Section 5. Experimental results are discussed in Section 6 followed by concluding remarks.

2. Curvelet based feature extraction

Fourier series decomposes a periodic function into sum of simple oscillating functions, i.e., sines and cosines. In a Fourier series sparsity is destroyed due to discontinuities and therefore numerous terms are required to precisely reconstruct a discontinuity. Multiresolution analysis tools were developed to overcome inherent limitations of Fourier series. Many fields of contemporary science and technology benefit from multiresolution analysis tools for maximum throughput, efficient resource utilization and accuracy. Multiresolution tools render robust behavior to study information content of images and signals in presence of noise and uncertainty.

Wavelet transform is a renowned multiresolution analysis tool that conveys accurate temporal and spatial information. Wavelet transform has been profusely used to address problems in data compression, pattern recognition and computer vision. Wavelets better represent objects with point singularities in 1D and 2D space but fail to deal with singularities along curves in 2D.

Discontinuities in 2D are spatially distributed which leads to extensive interaction between wavelet expansion coefficients. Therefore, wavelet representation does not offer sufficient sparseness for image analysis. In recent times, research community has witnessed intense efforts towards development of better directional and decomposition tools, namely, ridgelets [24] and contourlets [25].

Curvelet transform [26] is designed and targeted to represent smooth objects with discontinuity along a general curve. Curvelet transform overcomes shortcomings of existing multiresolution analysis schemes and offers improved directional capacity to represent edges and other singularities along curves. Curvelets are redundant bases that optimally represent 2D curves and outperform wavelets in situations that require optimal sparse representation of objects with edges, representation of wave propagators, image reconstruction with missing data, etc. In addition to scale and location, curvelet bases also capture information about orientation that fulfills parabolic anisotropic scaling law $width = length^2$ [26]. A curvelet is a combination of radial and angular window in frequency domain, defined in a polar coordinate system. This representation is constructed as a product of two windows, i.e., the angular and the radial dyadic frequential coronas. The angular window corresponds to a Radon transform (directional analysis) and the radial dyadic window emulates a bandpass filter whose cut off frequency extracts image information that follows a parabolic anisotropic scaling law [26]. Curvelet bases were designed to entirely cover the frequency domain, in contrast to other directional multiscale representations such as the Gabor transform [27] that result in a loss of information.

2.1. Continuous time curvelet transform

A number of algorithmic implementation strategies [28–30] based on curvelet's original architecture have been proposed. Lets consider a 2D space $\Re^2$, with a spatial variable x and a frequency-domain variable ω, and let r and θ represent polar coordinates in frequency-domain. $W(r)$ is used for radial partition, therefore its argument is positive. Whereas, $V(t)$ is used for angular partition and since the angle is assumed to take in both positive and negative values its argument is real. All derivatives of the windows exist and hence they are positive and compactly supported by arguments $r \in [1/2,1]$ and $t \in [-1,1]$. For every $j \geq j_0$, a frequency window U_j in the Fourier domain is defined as

$$U_j(r,\theta) = 2^{-3j/4} W(2^{-j}r) V\left(\frac{2^{\lfloor j/2 \rfloor}\theta}{2\pi}\right), \tag{1}$$

where $\lfloor j/2 \rfloor$ is the integral part of $j/2$. The support of U_j in polar coordinate system is a wedge (gray region in Fig. 3) defined by the support of W and V. A wedge spans an angle of $O(2^{(-j/2)})$ and covers a width of $O(2^{(j/2)})$. The windows W and V obey the following essential conditions in order to ensure a tight frame property:

$$\sum_{j=-\infty}^{+\infty} W^2(2^{-j}r) = 1, \quad r \in (3/4,3/2), \tag{2}$$

$$\sum_{l=-\infty}^{+\infty} V^2(t-l) = 1, \quad t \in (-1/2,1/2). \tag{3}$$

The window $W(r)$ covers the radial variable r in a multiscale manner. In Eq. (2), these windows are scaled by different powers of 2. When j approaches $-\infty$, the support of $W(2^j r)$, i.e., the set where the function is not zero goes to ∞, and when j approaches $+\infty$, the support goes to 0. Summation of j from $-\infty$ to $+\infty$ covers the range $[0,\infty]$. In this paper, r is selected as (3/4, 3/2), whereas, in general any interval such as $(\eta, 2*\eta)$ can be chosen, since the sum in j can be shifted with any integer shift without affecting the meaning of its summation. The window $V(t)$ covers the interval $[-1,1]$ and thus Eq. (3) ensures that the square of its integer shift adds to 1. The index l of Eq. (3) ranges from $-\infty$ to $+\infty$ because the argument of the function $V(\ )$ in Eq. (1) can be arbitrarily large. In Eq. (3), t spans from $(-1/2, 1/2)$, however, in principle any interval of length 1 can be selected.

Curvelets are defined (as function of $x=(x_1,x_2)$) at scale 2^{-j}, orientation θ_l, and position $x_k^{(j,l)} = R_{\theta_l}^{-1}(k_1 2^{-j}, k_2 2^{-j/2})$ by $\varphi_{j,k,l}(x) = \varphi_j(R_{\theta_l}(x - x_k^{(j,l)}))$, where R_θ is an orthogonal rotation matrix. A curvelet coefficient is evaluated by computing the inner product of an element $f \in L^2(R^2)$ and a curvelet $\varphi_{j,k,l}$:

$$c_{j,k,l} = \langle f, \varphi_{j,k,l} \rangle = \int_{R^2} f(x)\overline{\varphi_{j,k,l}}\, dx. \tag{4}$$

Curvelet transform also contains coarse scale elements similar to wavelet theory. For $k_1, k_2 \in Z$, we define a coarse level curvelet as

$$\varphi_{j_0,k}(x) = \varphi_{j_0,k}(x - 2^{-j_0}k), \quad \hat{\varphi}_{j_0(\omega)} = 2^{-j_0} W_0(2^{-j_0}|\omega|). \tag{5}$$

Curvelet transform is composed of fine-level directional elements $(\varphi_{j,k,l})_{j \geq j_0,k,l}$ and coarse-scale isotropic father wavelet $(\phi_{j_0,k})_k$. Key components of the construction are summarized in Fig. 3, left hand side represents the induced tiling of the Fourier frequency plane and the image on the right shows the associated spatial Cartesian grid at a given scale and orientation. The wedges are a consequence of Fourier plane partitioning in radial (concentric circles) and angular divisions. Concentric circles decompose the image in multiple scales (used for bandpassing the image) and angular divisions correspond to different angles or orientation. Therefore, to address a particular wedge we need to define both its scale and angle. Plancherel's theorem is applied in

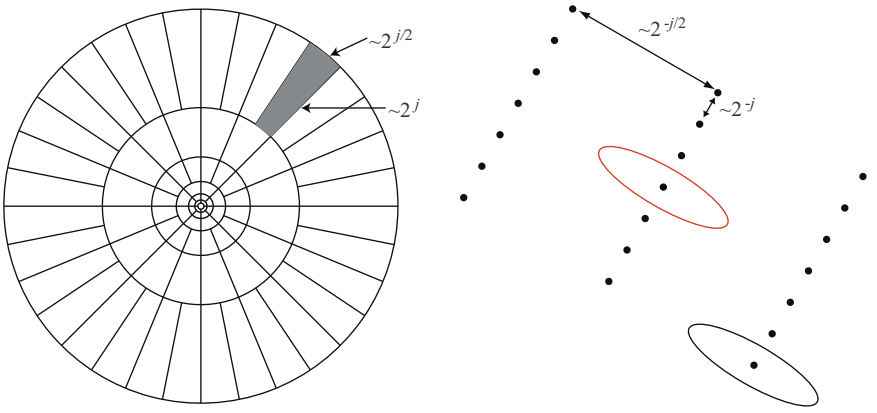

Fig. 3. Space–frequency tiling in Curvelet domain [26].

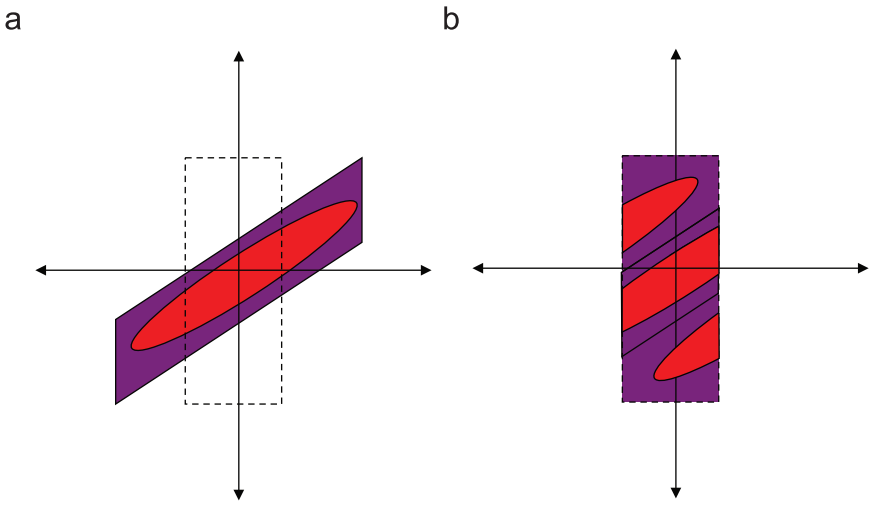

Fig. 4. Wrapping a segment around the origin [31]: (a) original and (b) wrapped.

Eq. (6) to express $c_{j,k,l}$ as an integral over the entire frequency plane.

$$c_{j,k,l} = \frac{1}{(2\pi)^2}\int \hat{f}(\omega)\overline{\hat{\varphi}_{j,k,l}(\omega)}\, d\omega = \frac{1}{(2\pi)^2}\int \hat{f}(\omega)U_j(R_{\theta_l}\omega)e^{i\langle x_k^{(j,l)},\omega\rangle}\, d\omega. \tag{6}$$

2.2. Fast discrete curvelet transform

New implementations of fast discrete curvelet transform (FDCT) are ideal for deployment in large-scale scientific applications due to their numerical isometry and an utmost 10 folds computational complexity as compared to FFT operating on a similar sized data. In our research work we used FDCT via wrapping [26] for image analysis.

- Compute 2D FFT coefficients and obtain Fourier samples $\hat{f}[n_1,n_2]$ where $-n/2 < n_1$ and $n_2 < n/2$.
- For each scale j and angle l, form the product $\tilde{U}_{j,l}[n_1,n_2]\hat{f}[n_1,n_2]$.
- Wrap this product around the origin and obtain $\tilde{f}_{j,l}[n_1,n_2] = W(\tilde{U}_{j,l}\hat{f})\ [n_1,n_2]$, where the range of n_1, n_2 and θ, respectively, are $0 < n_1 < L_{1,j}$, $0 < n_1 < L_{2,j}$ and $(-\pi/4,\pi/4)$.
- Apply inverse 2D FFT to each $\tilde{f}_{j,l}$ and save discrete curvelet coefficients.

In the first two stages Fourier frequency plane of the image is divided into radial and angular wedges owing to the parabolic relationship between a curvelet's length and width, as demonstrated in Fig. 3. Each wedge corresponds to curvelet coefficient at a particular scale and angle. Step 3 is essentially required to re-index the data around the origin as shown in Fig. 4. Finally, inverse FFT is applied to collect discrete curvelet coefficients in the spatial domain. Interested readers are requested to refer to [26] for additional mathematical details. In this work, a curvelet subband that demonstrates a maximum standard deviation is selected as an initial feature vector to represent each image. Dimensionality of feature vectors is reduced through application of our proposed B2DPCA algorithm.

3. Bidirectional two-dimensional principal component analysis

Karhunen–Loeve expansion, also known as principal component analysis (PCA), is a data representation technique widely used in pattern recognition and compression schemes. Pioneering work by Kirby and Sirovich [5] used PCA for enhanced representation of face images, however, PCA fails to capture minor variance unless they are explicitly accounted in the training data. Wiskott et al. [9] proposed a bunch graph matching technique to overcome limitations and flaws of linear PCA. Yang et al. [32], proposed two dimensional PCA for image representation. As opposed to PCA, 2DPCA is based on 2D matrices rather than 1D vectors. Therefore, image matrix does not need to be vectorized prior to feature extraction. Instead an image covariance matrix is directly computed using original image matrices.

Let X denote a q dimensional unitary column vector. To project a $p \times q$ image matrix A to X; linear transformation $Y=AX$ is used which results into a p dimensional projected vector Y. The total scatter of the projected samples is determined to measure the discriminatory power of the projection vector X. The total scatter is characterized by the trace of S_x, i.e., covariance matrix of the projected feature vectors; $J(X) = tr(S_x)$, where tr( ) represents the trace of S_x:

$$S_x = E[Y-E(Y)][Y-E(Y)]^T = E[(A-EA)X][(A-EA)X]^T, \tag{7}$$

$$tr(S_x) = X^T[E(A-EA)^T(A-EA)]X. \tag{8}$$

$G_t = E[(A-EA)^T(A-EA)]$ is a non-negative $q \times q$ image covariance matrix. If there are $\mathbb{M}$ training samples, the αth image sample is denoted by $p \times q$ matrix A_α:

$$G_t = \frac{1}{\mathbb{M}}\sum_{\alpha=1}^{\mathbb{M}}(A_\alpha-\check{A})^T(A_\alpha-\check{A}), \tag{9}$$

$$J(X) = X^T G_t X, \tag{10}$$

where $\check{A}$ represents an average image of all the training samples. The unitary vector X_{opt} that maximizes the generalized total scatter criterion $J(X)$ is called the optimal projection axes. X_{opt} represents a collection of M orthonormal eigenvectors $X_1,X_2,\ldots\ldots,X_M$ of G_t corresponding to M largest eigenvalues. Hence, dimensionality of every image A_α is reduced by post multiplying and pre-multiplying the image with optimal projection axes as $X_{opt}^T A_\alpha X_{opt}$.

A limitation of 2DPCA based recognition is its operability along row direction only. Zhang and Zhou [33] proposed $(2D)^2$ PCA based on an assumption that the training images are zero mean and thus, image covariance matrix can be computed using outer product of row/column image vectors. In [33], two image covariance matrices G_{tRow} and G_{tCol} are calculated by representing Eq. (9) initially in terms of row vectors of A_α and $\check{A}$, and repeating the similar operation for column vectors. The optimal projection axes of G_{tRow} and G_{tCol} are evaluated and labeled as X_{1opt} and Z_{1opt}.

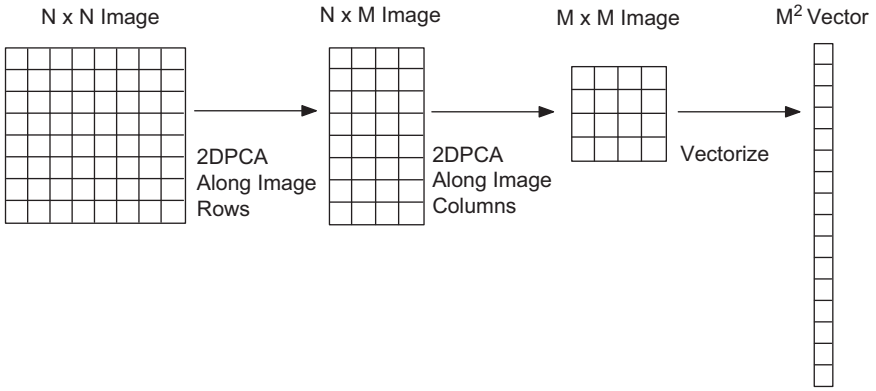

Fig. 5. Block diagram of proposed B2DPCA algorithm.

It is worth mentioning that both G_t and G_{tRow} are evaluated along rows and hence their projection axes X_{opt} and X_{1opt} are similar. A dimensionally reduced image of A_α is evaluated as $Z_{1opt}^T A_\alpha X_{1opt}$ [33].

Our dimensionality reduction algorithm operates independently along row and column directions (shown in Fig. 5) in order to better preserve the neighborhood relationship and to generate distinctive feature sets. Our proposed technique closely follows the work of [32] and generates an image covariance matrix $G_{t\alpha}$ and further optimizes it exploiting optimal project axes. Once optimal projection axes $X_{opt\alpha}$ is calculated, the dimensionality of every image A_α is reduced along its columns to generate new image sets A_β using Eq. (11). The newly generated image sets are subsequently treated as a fresh database and a latest image covariance matrix $G_{t\beta}$ and optimal projection axes $X_{opt\beta}$ are evaluated. Finally, every new image A_β is pre-multiplied by $X_{opt\beta}^T$ using Eq. (12). Hence, unlike traditional 2DPCA, a two fold approach is adopted in our proposed B2DPCA algorithm to reduce image dimensionality. We present the major implementation steps for computation of B2DPCA in Algorithm 1 for clarity and better readability:

$$A_\beta = A_\alpha X_{opt\alpha}, \tag{11}$$

$$A_\Theta = X_{opt\beta}^T A_\beta. \tag{12}$$

Algorithm 1. Proposed B2DPCA algorithm.

INPUT: Input images $A_{N\times N}$, $M \in \mathbb{Z}^+$
OUTPUT: $M \times M$ output matrix A_Θ
1: Compute non-negative image covariance matrix by $G_{t\alpha} = \frac{1}{\mathbb{M}} \sum_{\alpha=1}^{\mathbb{M}} (A_\alpha - \breve{A})^T (A_\alpha - \breve{A})$
2: The trace of the covariance matrix characterizes total scatter $J(X) = X^T G_{t\alpha} X$
3: $X_{opt\alpha} = \{X_1, X_2, \ldots, X_M\}$ where X_i represents a principal orthogonal vector
4: Reduce dimensionality along columns: $A_\beta = A_\alpha X_{opt\alpha}$
5: The image covariance matrix for A_β is determined as $G_{t\beta} = (1/\mathbb{M}) \sum_{\alpha=1}^{\mathbb{M}} (A_\beta - \breve{A}_\beta)^T (A_\beta - \breve{A}_\beta)$
6: Using total scatter criterion (step-2), compute $X_{opt\beta}$ comprising of M largest eigenvectors
7: Row-wise dimension reduction by $A_\Theta = X_{opt\beta}^T A_\beta$

4. Extreme learning machine

Feedforward neural networks are ideal classifiers for nonlinear mappings that utilize gradient descent approach for weights and bias optimization. The important factors that influence the performance of a traditional neural learning algorithm include:

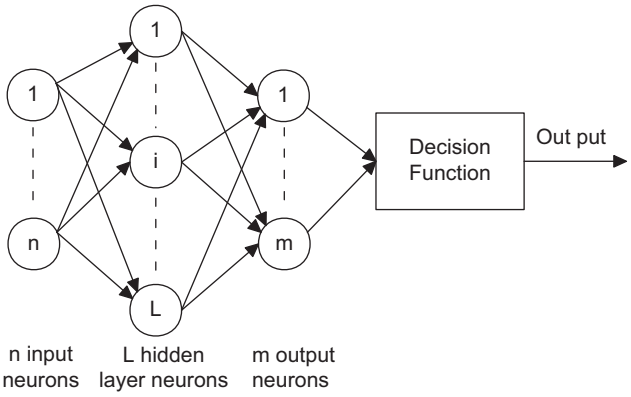

Fig. 6. Architecture of an extreme learning machine classifier.

- A small value of learning parameter $\breve{\rho}$ causes the learning algorithm to converge *slowly* whereas a higher value leads to *instability and divergence to a local minima.*
- Neural networks may be *over-trained* using back propagation (BP) and generate *inferior generalization performance.*
- Gradient descent based learning is an extremely *time consuming* process for most applications.

To overcome innate slow learning ability of traditional optimization techniques, Haung et al. [34] proposed ELM to train a single-hidden layer feedforward neural network (SLFNN) as shown in Fig. 6. A random selection of input weights and the hidden layer biases transforms the training of SLFNN into a linear system. Consequently, the output weights (linking the hidden layer and output layer) can be analytically determined through a simple generalized inverse operation of the hidden layer output matrices. In ELM, an infinitely differentiable activation function facilitates random assignment of input weights and hidden layer biases. Consider a collection of N distinct samples (x_i, t_i) where $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in \Re^n$ and $t_i = [t_{i1}, t_{i2}, \ldots, t_{im}]^T \in \Re^m$, an ELM with L hidden nodes and an activation function $\xi(x)$ is modeled as

$$\sum_{i=1}^{L} \gamma_i \xi_i(x_n) = \sum_{i=1}^{L} \gamma_i \xi_i(w_i x_n + b_i) = o_n, \quad n = 1,2,\ldots,N, \tag{13}$$

where $w_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$ and $\gamma_i = [\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iL}]^T$ represent input and hidden layer weight vectors, respectively. The ELM

reliably approximates N samples with minimum error:

$$\sum_{i=1}^{L} \gamma_i \xi_i(w_i x_n + b_i) = t_n, \quad n = 1,2,\ldots,N. \tag{14}$$

Eq. (14) can also be represented as $\delta\gamma = \tau$, $\delta = (w_1, \ldots, w_L, b_1, \ldots, b_L, x_1, \ldots, x_N)$, such that the ith column of δ is the output of the ith hidden node with respect to the inputs $x_1, x_2, \ldots, x_N$. If the activation function $\xi(x)$ is infinitely differentiable, it is proved that the number of hidden nodes are such that $L \ll N$. The training of ELM requires minimization of an error function E:

$$E = \sum_{n=1}^{N} \left(\sum_{i=1}^{L} \gamma_i \xi_i(w_i x_n + b_i) - t_n \right)^2 \Rightarrow E = \|\delta\gamma - \tau\|. \tag{15}$$

In classical neural networks δ is determined using gradient descent optimization wherein the input weights w_i, hidden layer weights γ_i and bias parameters b_i are iteratively tuned at a learning rate ρ. A small value of ρ causes the learning algorithm to converge slowly, whereas, a higher value leads to instability and divergence to a local minima. To avoid such instability and divergence to a local minima, ELM incorporates a minimum norm least-square solution. Therefore, instead of tuning the entire network parameters, input weights and bias parameters are randomly allocated and the problem is curtailed to a least-square solution of $\delta\gamma = \tau$. The hidden layer output matrix δ is a non-square matrix and the norm least-square solution reduces to $\gamma = \delta^*\tau$, where δ^* represents Moore–Penrose generalized inverse of δ. An infinitely small training error is achieved using ELM since it represents a least-square solution of the linear system:

$$\|\delta\hat{\gamma} - \tau\| = \|\delta\delta^*\tau - \tau\| \equiv \min_{\gamma}\|\delta\gamma - \tau\|. \tag{16}$$

Table 1
Mean standard deviation of curvelet subbands in various databases.

Database	Scale=1	Scale=2							
		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8
FERET	63.3	3.7	3.7	3.1	3.7	3.5	3.6	3.1	3.7
Faces94	74.6	4.9	4.3	3.6	3.9	5.0	4.5	3.4	3.4
JAFFE	87.2	6.7	6.9	3.7	4.4	6.7	6.9	3.7	4.3
GTech	68.9	3.0	2.8	4.9	4.6	3.3	3.3	5.5	5.4
ORL	56.1	5.1	3.9	6.1	5.9	3.9	3.5	5.4	5.3
Sheffield	51.0	5.2	3.4	4.0	5.1	7.8	4.4	3.6	4.8

5. Proposed face recognition algorithm

The proposed method is based on image decomposition of curvelet transform and uses dimensionally reduced coefficients for recognition. Distinctive feature sets generated using B2DPCA are used to train and test an ELM classifier. A block schematic diagram of our proposed algorithm is shown in Fig. 8.

Images from each database are converted into gray level image with a two fold reduction in image size. Each database is randomly divided into training and testing set so that 40–45% of images of each subject are used as prototypes and remaining images are used during testing phase. Curvelet transform is used to generate initial feature vectors since it offers superior performance in presence of singularities in higher dimension, and enhances localization of higher frequency components with minimized aliasing effects. Input images are resized to $R \times C$, since analogous image sizes support generation of curvelet feature vectors with identical level of global information. Furthermore, curvelet decomposition of all images within each database is computed at three scales and eight angular orientations thus, generating 25 distinct subbands.

The standard deviation of every subband is calculated and a subband that exhibits highest standard deviation is selected as an initial feature vector of size $U \times V$, where $U \times V \ll R \times C$. In contrast to the most recent work in literature [21] that uses two subbands, we have selected only one subband since the difference between standard deviations of the coarsest curvelet subband and the next coarser subband is quite significant. This noteworthy disparity in standard deviations is consistent for all the tested databases as shown in Table 1. The proposed approach is based on selecting a subband with the utmost standard deviation which leads to momentous savings in computational cost during dimensionality reduction. The curvelet transform of all images in every dataset is evaluated and standard deviation of all the curvelet subbands is determined. It is noticed that approximate curvelet subband holds the maximum standard deviation amongst all 25 curvelet subbands. Fig. 7 justifies our approach of selecting only one subband, i.e., curvelet subband at scale=1 and is in agreement with the results presented in Table 1.

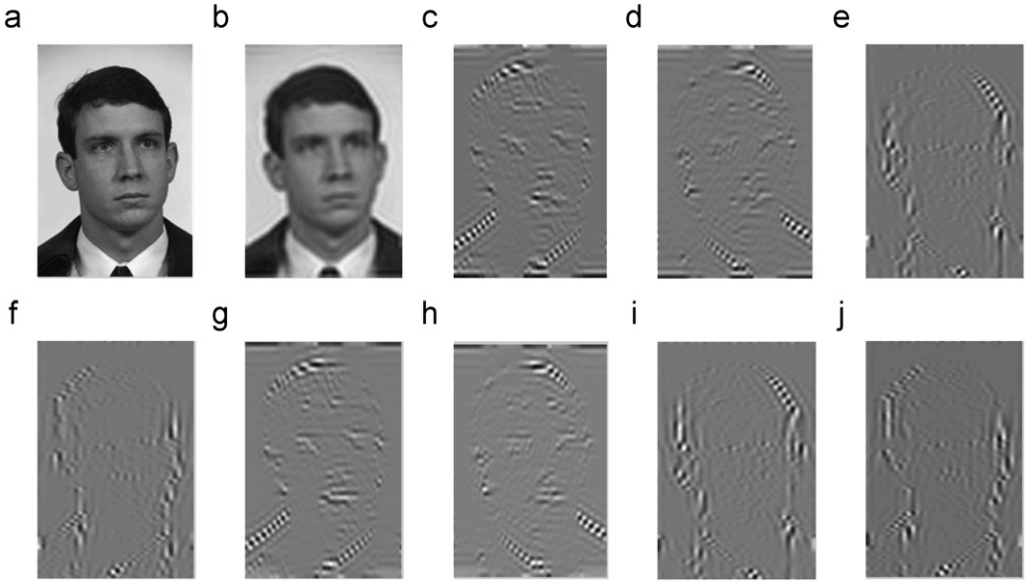

Fig. 7. (a) Original FERET image, (b) curvelet subband at scale=1, (c–j) curvelet subbands at scale=2 and 8 angular orientations.

B2DPCA is used to generate unique feature sets and to minimize computational complexity of our framework. Yang et al.'s [32] 2DPCA calculates a single covariance matrix to reduce the image dimensionality along its rows and columns, respectively, whereas, in our proposed approach image dimensionality along orthogonal directions is reduced independently of each axis. The intermediate features are extracted by initially reducing dimension of initial feature matrix, i.e., selected curvelet subband along its columns. Later dimensionality of intermediate features is reduced along its rows in order to generate a final feature set, each of size $U' \times V'$, where $U' \times V' \ll U \times V$ (refer to Section 3 for implementation details of B2DPCA). The modified approach used in this paper helps us to preserve critical neighborhood information between adjoining pixels and to generate distinctive features. For each dataset, dimensionally reduced curvelet feature sets are randomly selected for training of an ELM, whereas, remaining features of the same dataset are used to judge the separability of our framework. Note that we do not assume any *a priori* knowledge of the scene, background, face location, and illumination conditions. Our scheme belongs to the class of recognition techniques, which are based on global content representation without any requirement to locate the most probable location of a face in an image using automated face detection module or manual cropping to simplify the complicated task of recognition.

6. Experiments and results

Extensive experiments are performed using our proposed method on seven distinguishing face databases: FERET, Faces94, JAFFE [35], Georgia Tech [36], Sheffield [37], ORL [38] and YALE [6]. These are the most commonly used databases to evaluate and compare the performance of various face recognition algorithms. Before divulging into experimental details and results, we will briefly describe the databases used to vigorously test our algorithm.

6.1. Databases

The FERET database was sponsored by the Department of Defense in order to develop a system with automatic face recognition capability to be employed for assistance in security, intelligence and law enforcement. The final corpus consists of 14,051 eight-bit grayscale images of human heads with views ranging from frontal to left and right profiles, see [22] for more details.

Faces94 database was generated at the University of Essex and contains a series of 20 images per individual. Faces94 database is a wide-ranging database that contains images of 152 distinctive individuals. The database contains images of people of various racial origins, mainly first year undergraduate students, so the majority of individuals are between 18 and 20 years old. Older staff members and students are also included in the database where some individuals are wearing glasses and/or beards.

A Japanese female facial expression (JAFFE) database is also used to rigorously test the performance of our proposed method. The database contained 220 images of varying facial expressions posed by 10 Japanese female models.

Georgia Tech database contains images of 50 people and contains 15 color images for every subject. Most of the images are captured in two different sessions to take into account the variations in illumination conditions, facial expression, and appearance. Additionally images are acquired at varying scales and orientations.

Sheffield face Database consists of 564 images of 20 individuals. The database consist of images of individuals with mixed race, gender and appearance. Each individual is imaged in a range of poses from left/right profiles to frontal views with small angular rotations between successive images. The database has been pre-cropped so that the image size is uniformly reduced to 112×92, hence, background information is eliminated from input images and only the central characteristics of the face are retained.

ORL face database contains 10 different images for each of the 40 distinctive subjects. Subjects are imaged at different times, with varying lighting conditions, facial expressions and facial details. All images are captured against a dark homogeneous background with the subjects in an upright, frontal position with a small tolerance for side movement.

Yale face database contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and winking.

6.2. Comparative results

All Images are resized with a 2 fold dimension reduction and converted from RGB to gray level image. In all databases 40–45% of images of each subject are used as prototypes and the remaining images for testing purposes. Both the testing and training image sets are decomposed using curvelet transform at three scales and eight different angles. Approximate curvelet coefficients are dimensionally reduced using B2DPCA, vectorized, trained and tested using ELM. Fast learning and testing speed offered by ELM enabled us to repeat the experiments several times; every experiment is executed 100 times for each database and average results are reported.

Table 2
Comparative accuracy for YALE and ORL face database.

Method	YALE	ORL
Standard eigenface [5]	76	92.2
Waveletface [16]	83.3	92.5
Curveletface [21]	82.6	94.5
Waveletface + **PCA** [16]	84	94.5
Waveletface + **LDA** [17]	84.6	94.7
Waveletface + **weighted modular PCA** [19]	83.6	95
Curveletface + **LDA** [21]	83.5	95.6
Waveletface + **KAM** [18]	84	96.6
Curveletface + **PCA** [21]	83.9	96.6
Curveletface + **PCA** + **LDA** [21]	92	97.7
Curveletface + **B2DPCA** + **ELM**	99.7	99.9

Table 3
Average recognition rates (%) for Sheffield and FERET database.

Number of components	Sheffield		FERET	
	PCA+LDA	Proposed	PCA+LDA	Proposed
5	93.89	93.99	77.42	92.27
10	96.11	99.31	80.65	93.03
15	97.78	99.80	77.41	93.08
20	99.44	99.91	87.09	90.46
25	99.44	100	90.32	97.83
30	98.88	100	75.8	99.09
35	98.46	100	88.17	96.09
40	97.12	100	80.64	99.70
45	97.77	100	67.20	98.74
50	97.22	100	66.67	99.63

We have compared our ELM based recognition scheme (50 hidden neurons) against methods utilizing kNN of neighborhood size 5.

A comparative study of recognition performance of various techniques using ORL and Yale face databases is presented based on 60 principal components. It is evident from the results presented in Table 2 that our proposed method outperforms existing wavelet and/or curvelet based face recognition architectures. In the remaining experiments for simplicity we have only compared our results with a curvelet based PCA+LDA [21] approach.

The recognition accuracy achieved for Sheffield and FERET databases using varying number of principal components is compared with the curvelet based PCA+LDA approach [21] in Table 3, whereas, results obtained for ORL and GTech databases are listed in Table 4. The recognition achieved using our proposed method consistently outperform PCA+LDA based approach for Sheffield, FERET, ORL and GTech datasets. The improvements in accuracy are mainly pronounced for FERET database making it obvious that our method is suitable to deal with challenging databases (views ranging from front to left and right profiles at varying orientations). It is worth mentioning that increasing the number of principal components does not necessarily increase the accuracy and the use of localized information for face recognition may be exploited to generate improved results.

Table 4
Average recognition rates (%) for ORL and GTech database.

Number of components	ORL		GTech	
	PCA+LDA	Proposed	PCA+LDA	Proposed
5	79.12	94.05	88.32	89.14
10	89.16	99.56	71	93.53
15	94.21	98.19	90.33	97.43
20	98.33	99.73	95.65	97.09
25	97.5	99.56	96.34	96.81
30	97.5	99.94	94.67	97
35	98.42	99.96	96	97.42
40	96.67	99.99	96	97.71
45	97.45	100	94	97.6
50	97.52	100	93	97.87

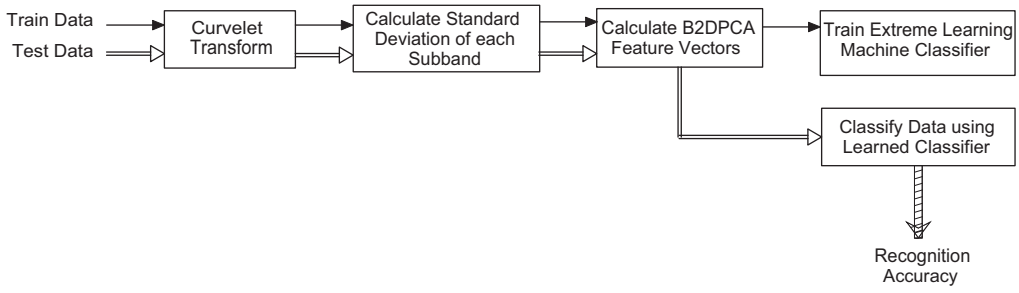

Fig. 8. Schematic diagram of proposed face recognition algorithm.

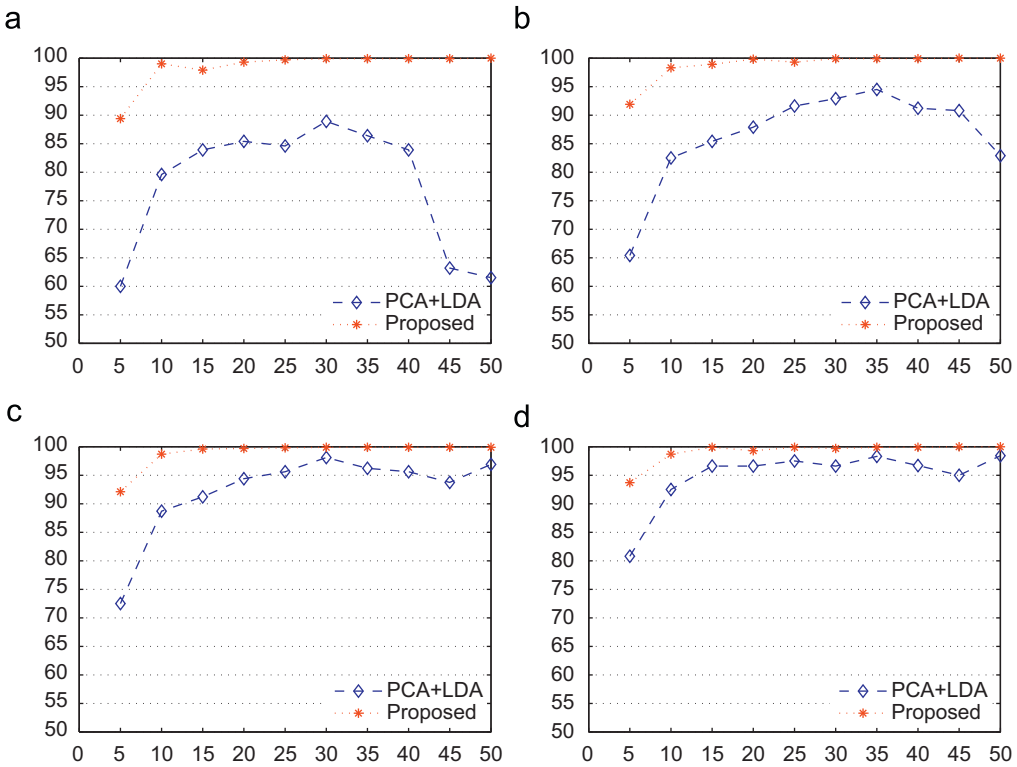

Fig. 9. Average recognition rate (*y*-axis) vs. number of principal components (*x*-axis) for ORL database at varying prototypes. (a) 30% prototypes, (b) 40% Prototypes, (c) 60% prototypes and (d) 70% Prototypes.

In addition to improved accuracy, our proposed method is also independent of the number of prototypes in comparison to other face recognition algorithms. Recognition rates obtained for ORL database at 30%, 40%, 60% and 70% prototypes are plotted in Fig. 9 (y-axis denotes the accuracy and x-axis denotes the number of principal components). In order to avoid within-scatter matrix singular cases, authors in [21] extracted curvelet coefficients at four scales. In contrast, our proposed method is robust and free of the singularity issues, i.e., independent of the scales of curvelet decomposition that radically degrade precision of the PCA+LDA based method.

Table 5 compares average recognition rates (AVR) and time complexity for Faces94 database. Results clearly validate our claim that the proposed method achieves superior recognition at hundred folds faster speed than state-of-the-art technique [21], and is suitable for real-time applications. In addition to improvements in classification time, our system also achieves significant savings in computational time during dimensionality reduction stage since only one subband is utilized as a feature matrix.

In order to emphasize the benefits of our proposed dimensionality reduction technique, i.e., B2DPCA, we compared the accuracy achieved using Yang's 2DPCA [32] against our approach. In both situations we used an ELM classifier to train our system and to ascertain recognition rate. In Table 6, average recognition rates obtained for FERET database are compared for varying principal components. It is worth mentioning that the number of principal components are represented in the form of square of integers because of operational behavior of 2DPCA that simultaneously reduces dimensions along rows and columns using a single set of optimized eigenvectors (refer to Section 3 for details). The improvements in accuracy using our proposed dimensionality reduction technique are apparent from the results.

Table 5
Average recognition rates (%) and time complexity for Faces94 database.

Number of components	PC+LDA		Proposed	
	AVR (%)	Time (s)	AVR (%)	Time (s)
10	92.17	30.74	94.92	0.1329
20	97.28	30.67	98.91	0.1340
30	99.29	31.55	99.54	0.1343
40	99.29	33.17	99.55	0.1321
50	99.29	33.36	99.87	0.1348

Table 6
Average recognition rates (%) for FERET database using 2DPCA and B2DPCA.

Number of components	2DPCA+ELM	B2DPCA+ELM
	AVR (%)	AVR (%)
4	49.7	51.06
9	83.12	77.93
16	78.15	93.91
25	94.13	97.83
36	99.25	99.74
49	99.78	99.63

Table 7
Average recognition rates (%) for JAFFE database at varying number of neurons.

	Neurons						
Components	35	40	45	50	55	60	STD
5	92.56	92.92	92.97	92.95	92.62	92.55	0.2047
10	99.80	99.78	99.93	99.77	99.81	99.75	0.0641
15	99.01	99.07	99.01	98.98	99.04	98.94	0.0454
20	99.97	99.95	99.96	99.97	99.89	99.95	0.0299
25	100	100	100	100	100	100	0

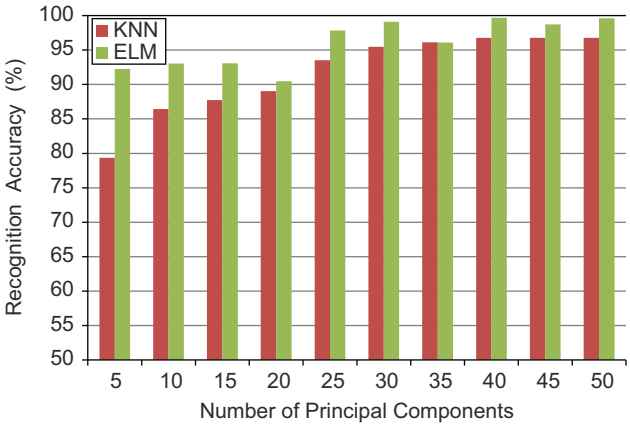

Fig. 10. Average recognition rate (%) for FERET database using kNN and ELM.

Experiments are also carried out by varying the number of hidden neurons from 35 to 60 in intervals of 5, however, negligible variations in accuracy are observed, as indicated by the recognition rates and standard deviation (STD) in Table 7. The results represent a significantly improved behavior as compared to traditional classification schemes whose correctness is greatly attributed to various parameters, for example, neighborhood size for a kNN classifier. To further emphasize the advantages associated with the use of an ELM classifier, we classified B2DPCA reduced feature vectors using a kNN and ELM classifier with five neighbors and 50 hidden neurons, respectively. Improved recognition accuracy is achieved using ELM in comparison with kNN at varying number of principal components, as presented in Fig. 10.

7. Conclusion

In this paper an efficient human face recognition technique based on curvelet feature subspace is proposed. The curvelet transform is used to compute sparse features with improved directionality in higher dimension. These sparse features are dimensionally reduced using B2DPCA to generate distinctive feature sets. Finally, these features are input to an ELM to analytically learn an optimal model. Experimental results corroborate our claim that the proposed method achieves improved recognition at a substantially faster rate against existing techniques. In addition, our proposed method is independent of the number of prototypes used for training, scales of curvelet decomposition and the number of hidden neurons. In future, we would like to explore the use of localized features integrated with the curvelet based global information on recognition accuracy and classification speed. Law enforcement, border security, video surveillance and database security applications can potentially benefit from our proposed recognition scheme.

Acknowledgments

The work is supported in part by the Canada Research Chair Program and the NSERC Discovery Grant. Authors are thankful to curvelet.org team for useful links and would also like to express their gratitude to all researchers who provided image databases.

References

[1] L.F. Chen, H.Y. Mark Liao, C.C. Han, J.C. Lin, Why recognition in a statistics-based face recognition system should be based on the pure face portion:

a probabilistic decision-based proof, Pattern Recognition 34 (5) (2001) 1393–1403.
[2] B.S. Manjunath, R. Chellappa, C.V. Malsburg, A feature based approach to face recognition, International Conference on Computer Vision and Pattern Recognition (1992) 373–378.
[3] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, N. Otsu, Face recognition system using local autocorrelations and multiscale integration, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (10) (1996) 1024–1028.
[4] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, Face recognition: a literature survey, ACM Computing Surveys 35 (4) (2003) 399–458.
[5] M. Kirby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1) (1990) 103–108.
[6] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face recognition by independent component analysis, IEEE Transactions on Neural Networks 13 (6) (2002) 1450–1464.
[7] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using LDA-based algorithms, IEEE Transactions on Neural Networks 14 (1) (2003) 195–200.
[8] C. Liu, H. Wechsler, Evolutionary pursuit and its application to face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (6) (2000) 570–582.
[9] L. Wiskott, J.M. Fellus, N. Kruger, C. VonDerMalsburg, Face recognition by elastic bunch graph matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 775–779.
[10] F.R. Bach, M.I. Jordan, Kernel independent component analysis, Journal of Machine Learning Research 3 (2002) 1–48.
[11] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, IEEE Transactions on Neural Networks 14 (1) (2003) 117–126.
[12] C. Liu, H. Wechsler, A unified Bayesian framework for face recognition, International Conference on Image Processing (1998) 151–155.
[13] B. Moghaddam, T. Jebara, A. Pentland, Bayesian face recognition, Pattern Recognition 33 (11) (2000) 1771–1782.
[14] K. Jonsson, J. Matas, J. Kittler, Y.P. Li, Learning support vectors for face verification and recognition, International Conference on Automatic Face and Gesture Recognition (2000) 208–213.
[15] B. Heisele, P. Ho, T. Poggio, Face recognition with support vector machines: global versus component-based approach, International Conference on Computer Vision 2 (2001) 688–694.
[16] G.C. Feng, P.C. Yuen, D.Q. Dai, Human face recognition using PCA on wavelet subband, Journal of Electronic Imaging 9 (2) (2000) 226–233.
[17] J.T. Chien, C.C. Wu, Discriminant waveletfaces and nearest feature classifiers for face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2) (2002) 1644–1649.
[18] B.L. Zhang, H. Zhang, S. Sam Ge, Face recognition by applying wavelet subband representation and kernel associative memory, IEEE Transactions on Neural Networks 15 (1) (2004) 166–177.
[19] M. Zhao, P. Li, Z. Liu, Face recognition based on wavelet transform weighted modular PCA, Proceedings of the Congress in Image and Signal Processing (2008) 589–593.
[20] D.L. Donoho, M.R. Duncan, Digital curvelet transform: strategy, implementation and experiments, Proceedings of SPIE 4056 (2000) 12–30.
[21] T. Mandal, Q.M.J. Wu, Y. Yuan, Curvelet based face recognition via dimension reduction, Elsevier Signal Processing 89 (3) (2009) 2345–2353.
[22] P.J. Phillips, H. Wechsler, J. Huang, P.J. Rauss, The FERET database and evaluation procedure for face-recognition algorithms, Elsevier Image and Vision Computing 16 (5) (1998) 295–306.
[23] L. Spacek, The essex faces94 database 〈http://cswww.essex.ac.uk/mv/allfaces/〉.
[24] M.N. Do, M. Vetterli, The finite ridgelet transform for image representation, IEEE Transactions on Image Processing 12 (1) (2003) 16–28.
[25] M.N. Do, M. Vetterli, The contourlet transform: an efficient directional multiresolution image representation, IEEE Transactions on Image Processing 14 (12) (2005) 2091–2106.
[26] E.J. Candes, L. Demanet, D.L. Donoho, L. Ying, Fast discrete curvelet transforms, Multiscale Modeling and Simulation 5 (3) (2006) 861–899.
[27] E.J. Candes, F. Guo, New multiscale transforms, minimum total variation synthesis: applications to edge-preserving image reconstruction, Signal Processing: Applications to Edge 82 (11) (2002) 1519–1543.
[28] J.L. Starck, N. Aghanim, O. Forni, Detecting cosmological non-Gaussian signatures by multi-scale methods, Astronomy and Astrophysics 416 (1) (2004) 9–17.
[29] J.L. Starck, M. Elad, D.L. Donoho, Redundant multiscale transforms and their application for morphological component analysis, Advances in Imaging and Electron Physics 132 (2004) 287–342.
[30] F.J. Herrmann, U. Boniger, D.J. Verschuur, Nonlinear primary-multiple separation with directional curvelet frames, Geophysical International Journal 170 (2) (2007) 781–799.
[31] B. Eriksson 〈www.homepages.cae.wisc.edu/∼ece734/∼project/∼s06/∼eriksson.ppt〉.
[32] J. Yang, D. Zhang, A.F. Frangi, J. Yang, Two-dimensional PCA: a new approach to appearance based face representation and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (1) (2004) 131–137.
[33] D. Zhang, Z.H. Zhou, (2D)2 PCA: two-directional two-dimensional PCA for efficient face representation and recognition, Elsevier Neurocomputing 69 (1) (2005) 224–231.
[34] G. Huang, Q. Zhu, C. Siew, Extreme learning machine: theory and applications, Elsevier Neurocomputing 70 (1–3) (2006) 489–501.
[35] M.J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (12) (1999) 1357–1362.
[36] A.V. Nefian, M. Khosravi, M.H. Hayes, Real-time human face detection from uncontrolled environments, SPIE Visual Communications on Image Processing (1997).
[37] D.B. Graham, N.M. Allinson, Characterizing virtual eigensignatures for general purpose face recognition, NATO ASI Series F, Computer and Systems Sciences: From Theory to Applications 163 (1998) 446–456.
[38] F. Samaria, A. Harter, Parameterisation of a stochastic model for human face identification, Second IEEE Workshop on Applications of Computer Vision (1994) 138–142.

Abdul Adeel Mohammed is a Post Doctoral Fellow at the University of Waterloo, Canada. He received his Ph.D. degree in Electrical Engineering from the University of Windsor, Canada in 2010. He completed his B.E. in 2001 from Osmania University, India, M.A.Sc. in 2005 from Ryerson University, Canada. His main area of research is 3D pose estimation, robotics, computer vision, sensor fusion, image compression and coding theory.

Rashid Minhas is working with Computer Vision and Sensing Systems Laboratory, University of Windsor (UWindsor), Canada. He also received his Ph.D. degree in Electrical Engineering (2010) from UWindsor, Canada. He is the recipient of IITA Scholarship (Korea), Fredrick Atkins Graduate Award (UWindsor), and Ministry of Research and Innovation—Post-Doctoral Fellowship, ON, Canada. His research interests include object and action recognition, shape reconstruction and fusion using machine learning and statistical techniques.

Q.M. Jonathan Wu (M'92, SM'09) received his Ph.D. degree in Electrical Engineering from the University of Wales, Swansea, UK, in 1990. From 1995, he worked at the National Research Council of Canada (NRC) for 10 years where he became a Senior Research Officer and Group Leader. He is currently a Professor in the Department of Electrical and Computer Engineering at the University of Windsor, Canada. Dr. Wu holds the Tier1 Canada Research Chair (CRC) in Automotive Sensors and Sensing Systems.He has published more than 150 peer-reviewed papers in areas of computer vision, image processing, intelligent systems, robotics, micro-sensors and actuators, and integrated micro-systems. His current research interests include 3D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks. Dr. Wu is an Associate Editor for IEEE Transaction on Systems, Man, and Cybernetics (Part A). Dr. Wu has served on the Technical Program Committees and International Advisory Committees for many prestigious conferences.

Maher A. Sid-Ahmed received Ph.D. from the Department of Electrical and Computer Engineering at University of Windsor ON Canada in 1974. He is a Professor and Chair of the Department of Electrical and Computer Engineering, University of Windsor Canada. Dr. Sid-Ahmed holds four U.S. patents in the area of real time video processing and improved definition Television and has authored over 100 papers in the areas of DSP, Improved Definition Television, OCR, Machine Vision and Metrology, Architectural design and VLSI.