3

A Multi-Stage Classifier for Face Recognition Undertaken by Coarse-to-fine Strategy

Jiann-Der Lee[1] and Chen-Hui Kuo[2]
[1]*Department of Electrical Engineering, Chang Gung University, Tao-Yuan, 333*
[2]*Department of Electrical Engineering, Chung Chou Institute of Technology, Chang-Hua, 510, Taiwan , R.O.C*

1. Introduction

Face recognition has been a very active research area for past two decades due to its widely applications such as identity authentication, airport security and access control, surveillance, and video retrieval systems, etc. Numerous approaches have been proposed for face recognition and considerable successes have been reported [1]. A successful face recognition system should be robust under a variety of conditions, such as varying illuminations, pose, expression, and backgrounds. Although many researches [1-3] already found solutions to alleviate those problems, it is still a great challenge to accurately recognize faces that are non-frontal facial images, and disguises such as facial hair, glasses or cosmetics.

In the past, two categories of feature extraction for face recognition have been developed, including of geometric features and statistical features derived from face images [1-4]. In geometric features method, the facial features are retrieved from either the shapes of eyes, nose, mouth, and chin, or the facial geometrical relationships such as areas, distances, and angles [4]. This kind of approaches has proven to be difficult for practical application because it requires a fine segmentation of facial features. In statistical features method, the facial features are usually extracted from important facial features based on the high-dimensional intensity values of face images. For example, the principal component analysis (PCA) is the well-known statistical method [5]. This approach is simple, but does not reflect the details of facial local features.

Generally speaking, researches on face recognition system can be grouped into two categories of classifier system, one is single-classifier system and the other is multi-classifier system. The single-classifier systems, including neural network (NN) [6], Eigenface [5], Fisher linear discriminant (FLD) [7], SVM [8], HMM [9], and AdaBoost [10], are developed well in theory and experiment. On the other hand, the multi-classifier systems such as local and global face information fusion [11-13], neural networks committee (NNC) [14], multi-classifier system (MCS) [15], are proposed in parallel process of different features or classifiers.

As one knows, neural networks (NN) are a nonlinear classifier and based on the parallel architecture of human brains. Its output can be binary or multiple classes. In NN, the margin between two classes of sample point nearby the decision function is often not a maximum.

Source: State of the Art in Face Recognition, Book edited by: Dr. Mario I. Chacon M.,
ISBN -3-902613-42-4, pp. 250, January 2009, I-Tech, Vienna, Austria

Because of its performance surface depending on the mean square error, and if the train error reaches to zero too quickly, its performance surface will be minimum and the weight values will have to be stopped to get adjusted [16]. For each subject, the numbers of face image sample are always small on public face databases, so the early training break in NN frequently occurs. The AdaBoost also has the same drawback as the NN does. That is, the situation of the early training break occurs in AdaBoost if the error of weak learner quickly reaches to zero. Besides, the application of HMM is a content-dependent classification, especially on speech recognition. It is based on an assumption that the class variations are closely related. The research of HMM on face recognition includes 1D-HMM [17], 2D-HMM [18], embedded HMM [19], and pseudo 2D-HMM [20]. The disadvantage of using HMM on face recognition is on the part of huge computation cost under the condition that the numbers of dimension and state are high.

The SVM were originally designed for binary classification and it is based on the structural risk minimization (SRM) principle. Although several methods to effectively extend the SVM for multi-class classification have been reported on technical literatures [21,22], it is still an on-going research issue. The category methods of SVM for multi-class classification are one-against-all (OAA) [21,30], one-against-one (OAO) [21], directed acyclic graph support vector machine (DAGSVM) [23], and binary tree SVM [8]. In DAGSVM and binary tree SVM, their training phases are the same as those of the OAO method by solving $N(N\text{-}1)/2$ binary SVMs, where N is the numbers of class. However, in the testing phase of DAGSVM, it uses a rooted binary directed acyclic graph which has $N(N\text{-}1)/2$ internal nodes and N leaves. Each node is a binary SVM of ith and jth classes. On the other hand, the testing phase of binary tree SVM constructs a bottom-up binary tree for classification. The advantage of using a DAGSVM and binary tree SVM is less testing time than that of OAO (maximum vote) method. If one employs the same feature vector for SVM, NN, and AdaBoost, he will find the performance of SVM is better than that of NN and AdaBoost because the SVM will result in the maximum separating margin to the hyperplane of the two classes. And if the feature vector includes noisy data, and the noisy data possesses at least one of the following properties: (a) overlapping class probability distributions, (b) outliers and (c) mislabeled patterns [24], the hyperplane of SVM will turn out to be hard margin and overfitting. Additionally, the SVM allows noise or imperfect separation, provided that a non-negative slack variable is added to the objective function as a penalizing term.

The Eigenface is a successful example using template matching for face recognition [5]. The method of Eigenface uses principle component analysis (PCA) or Karhunen-Loeve transform that constructs a number of Eigenfaces derived from a set of training face images. Every prototype face image in the database is represented as a feature point, i.e. a vector of weights, in the space and is the query face image. The limitations of Eigenface are their computation cost and memory requirement burden if the face recognition system is to scale up.

As for multi-classifier systems, Zhou et al. [13] presented a combined feature Fisher classifier (CF^{2}C) approach, whose combined facial features are derived from facial global and local information extracted by DCT, for face recognition. And it performs better than the traditional methods such as Eigenfaces and Fisherfaces. One problem of this method is that it is hard to accurately detect the landmark of images, and another is that the localization errors might sustain to the classification step to disturb the final result. Rajagopalan et al. [12] proposed a face recognition method that fuses information acquired from global and local features of the face for improving performance. Their method is very similar to that proposed by Zhou et al. [13], it brings about the same problem accordingly as mentioned above. Kwak et al. [11] proposed two approaches to fuzzy information fusion for face recognition involving aggregation of local and global face information and a wavelet decomposition approach. Zhao et al. [14] studied a face recognition method based on multi-features using a neural networks committee (NNC) machine. Either of their methods is based on parallel features and classifiers, and is with a combination strategy at the end. Although their experimental results show a more accurate classification rate than that of single feature and classifier, its architecture is totally different from our proposed cascade stages, which are proceeded with a coarse-to-fine strategy. Our novel coarse-to-fine strategy ends up with a much more successful performance in facial recognition.

In our system, the extracted feature for SVM is discrete cosine transform (DCT) coefficients that are common used for image pre-processing. In the past, Chen et al. [6] extracted DCT feature from the entire face image for face recognition, because they believed that if the DCT is obtained from individual sub-images, certain relationship information between sub-images are not existed any more. In [25], Jing and Zhang selected the useful DCT frequency bands and obtained a 1D training sample set, then they proposed an improved Fisherface method to extract the image discrimination features, at last they applied the nearest neighbor classifier to the feature classification.

To combine the image feature of frequency, intensity, and space information, we propose a novel face recognition approach, which combines SVM, Eigenface, and RANSAC [26] methods with the multi-stage classifier system. The whole decision process is developed through consecutive stages, i. e., “one-against-all (OAA) of SVM”, “one-against-one (OAO) of SVM”, “Eigenface”, and “RANSAC”, respectively. The stage 1 “OAA of SVM” and stage 2 “OAO of SVM” uses the DCT features extracted from the entire face image. The stage 3 “Eigenface”, face images are projected onto a feature space (face space). The face space is defined by the “Eigenfaces”, which are the eigenvectors of the set of faces and based on intensity information of the face image. “RANSAC” is applied in the last stage, in which the epipolar geometry method with space information of the testing image is matched with the two training images, and then the image with the greatest match numbers of and the shortest distance to corresponding feature points is selected. The face databases used for performance evaluation are retrieved from Olivetti Research Laboratory (ORL) [44], Yale database [45], and IIS (Institute of Information Science, Academia Sinica) face databases [46]. For each database, we use four different evaluation methods, which are OAA-SVM, OAO-SVM, Eigenface, and our proposed multi-stage classifier. In this way, the experimental results can be compared with other face recognition approaches fairly.

The remainder of this paper is organized as following: In section 2, the feature extraction methods, i.e., the DCT and PCA are briefly described. In section 3, the proposed coarse-to-fine stages, OAA, OAO, and multi-stage classifiers are presented in detail. Experimental results using this method and the comparison to other approaches with several famous face databases are given in section 4. Conclusions are included in section 5.

2. Feature extraction

The goal of feature extraction is to transform a given set of sample data to a new set of features. If the transform method is suitably chosen, the features after transformation can exhibit high "information packing" properties compared with the original input data. This means that most of the classification-related information is "squeezed" in a relatively small number of features, leading to a reduction of the necessary feature space dimension. 2D image that usually transforms space information to frequency information can be accomplished by one of the following methods, such as Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Fast Fourier Transform (FFT). These methods are applicable to fixed original images and optimal information packing properties. In addition, the computation requirement of these methods is lower than that of Karhunen-Loeve (KL, also known as PCA) Transform [6]. PCA is a useful statistical technique that can be applied to fields such as face recognition [5] and image compression, and it is a common technique for finding patterns in data of high dimension. PCA also is for feature extraction under the unsupervised learning setting. This section will give a brief introduction of the two methods, DCT and Eigenface, for feature extraction.

2.1 Discrete Cosine Transform (DCT)

There are three procedures for feature extraction of frequency transformation: (1). data sampling, (2) data transform, (3) feature vector extraction. For data sampling, two kinds of approaches namely block-based and non block-based are common used; overlapped block-based with top-bottom scan [27] or overlapped block-based with raster scan [28,29], and non block-based Full image [6]. That is, in this approach, we employ the DCT to the entire face image for data sampling, as shown in Fig. 1. Because some related information between sub-images cannot be obtained if the DCT is applied to the sub-image independently. With entire-face data sampling by the DCT, all frequency information then can be accomplished. The DCT coefficients with large magnitude are mainly located in the upper-left corner of the DCT matrix as shown in Fig. 1(c). In our system, we scan the DCT coefficient matrix in a zig-zag manner starting from the upper-left corner and subsequently convert it to a one-dimensional vector as shown in Fig. 1(d).

The DCT is a technique used for converting space signals into frequency components. The one-dimensional DCT is useful in processing one-dimensional (1D) signals such as speech waveforms. For analysis of two-dimensional (2D) signals such as images, a 2D version of the DCT (2D-DCT) is required. In short, for an $N \times M$ matrix f, the 2D-DCT is computed in a simple way: The 1D-DCT is applied to each row of f and then to each column of the result. If an image $f(x,y)$ of size $N \times M$, whose discrete cosine transform is $C(u,v)$, it can be expressed as Eq. (1).

$$C(u,v) = \frac{2}{\sqrt{NM}} \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} f(x,y) \cos\left(\frac{(2x+1)u\pi}{2N}\right) \cos\left(\frac{(2x+1)v\pi}{2M}\right) , \qquad (1)$$

$$u = 0, \ldots, N, \qquad v = 0, \ldots, M$$

where $$\alpha(u) = \begin{cases} 2^{-1/2} & for \quad u = 0 \\ 1 & otherwise \end{cases}$$

Since 2D-DCT can be computed by applying 1D transforms separately to the rows and columns, it can be said that 2D-DCT is separable in two dimensions. In the 1D case, each element $C(u,v)$ of the transform is the inner product of the input and a basis function, but in the 2D case, the basis functions are $N \times M$ matrices. Each 2D basis matrix is the outer product of two of the one-dimensional basis vectors.

2.2 Eigenface method

PCA is a well-known technique commonly exploited in multivariate linear data analysis. The main underlying concept is to reduce the dimensionality of a data set while retaining as much variation as possible in a data set. If a face image $f(x,y)$ is a two-dimensional $N \times N$ array of intensity values, the corresponding image i_k is viewed as a vector with N^2 coordinates that result from a concatenation of successive rows of the image. Moreover, if the training sets have M face images and every image is of the same size, the training set of M faces can be denoted by $I = \{i_1, i_2, \ldots, i_M\}$. Its average face $\bar{i}$ is defined by:

$$\bar{i} = \frac{1}{M}\sum_{j=1}^{M} i_j$$

Each face image differs from the average by the vector $\phi_j = i_j - \bar{i}$ for j=1,…, M. This set of huge vectors is then subject to principal component analysis, which seeks a set of M orthonormal vectors u_k and their associated eigenvalues λ_k , which best describes the distribution of the data. The vectors u_k and scalars λ_k are the eigenvectors and eigenvalues, respectively, of the covariance matrix.

$$\begin{aligned} C &= \frac{1}{M}\sum_{j=1}^{M} \phi_j \phi_j^T \\ &= AA^T \end{aligned} \tag{2}$$

where $A=[\phi_1\ \phi_2\ \ldots \phi_M]$. The matrix C of Eq. (2), however, is $N^2 \times N^2$, and determining the N^2 eigenvectors and eigenvalues is an intractable task for typical image sizes. This problem can be alleviated by referring to some basic finding known in linear algebra. Fortunately we can determine the eigenvectors by first solving a much smaller $M \times M$ matrix problem, and taking linear combinations of resulting vectors [5]. The cost of computation is greatly decreased from the order of the number of pixels in the images (N^2) to the order of the number of images in the training set (M). In practice, the numbers of face images M are smaller than numbers of pixels N^2 ($N^2 >> M$), and the calculations of the covariance matrix C become quite manageable.

A testing face image (i_x) is transformed into its eigenface components (projected into "face space") by a simple operation, $w_k = u_k^T(i_x - \bar{i})$ for k=1,…, M. The weights form a vector $\psi^T = [w_1\ w_2\ \ldots\ w_M]$ that describes the contribution of each eigenface in representing the input face image, treating the eigenfaces as a basis set for face images. The vector is used to find which number of pre-defined face class best describes the face. The simplest method for determining which face class provides the best description of an input face image is to find the face class k that minimizes the Euclidian distance $\varepsilon_k = \left\|(\psi - \psi_k)\right\|$, where ψ_k is a vector describing the kth face class.

3. The proposed method

For face recognition, based on coarse-to-fine strategy, we design a multi-stage recognition system which combines SVM, Eigenface, and RANSAC methods to increase the recognition accuracy. The detail of this system is demonstrated as following.

3.1 Support vector machine for binary classifier

Support vector machine (SVM) is a method based on statistical learning theory, which is applicable as pattern recognition, developed by V Vapnik [22]. The main purpose of the SVM is to separate two classes and maximize the margin between Hyper-plane and the nearly data point, as shown in Fig. 2.

In the case of linearly separating two classes, the two classes of hyper-planes, $(w{\cdot}x)+b=0$, where $w \in R^d$ and $b \in R$, is considered, its corresponding to the decision function Eq. (3) is:

$$f(x) = \operatorname{sgn}((w \cdot x) + b) = \pm 1 \tag{3}$$

The decision function $f(x)$ is described by weight vector w, threshold b and input patterns x. The solution to the optimization problem of SVM is given by the saddle point of Lagrange functional as Eq. (4):

$$\begin{cases} \min\limits_{w,b} L_p = \dfrac{1}{2}\|w\|^2 - \sum\limits_{i=1}^{m} \alpha_i y_i \cdot ((x_i \cdot w) + b) + \sum\limits_{i=1}^{n} \alpha_i \\ subject\ \ to \quad \alpha_i \geq 0 \quad i = 1, 2, \ldots, n \\ L_p : primal \quad problem \end{cases} \tag{4}$$

where α_i : Lagrange multipliers.

By using Lagrange multiplier techniques, the minimization of Eq. (4) leads to the following dual optimization problem as Eq. (5).

$$\begin{cases} \max\limits_{\alpha_i} \quad L_D = \sum\limits_{i=1}^{n} \alpha_i - \dfrac{1}{2} \sum\limits_{i=1;j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j \\ subject\ to \quad \alpha_i \geq 0 \quad i = 1, 2, \ldots, n \\ \qquad\qquad \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{5}$$

where L_D: dual problem

When the training data are nonlinear, we usually change the feature space dimension using the transfer function Φ( ). Thus the decision functions of the more general form [31] are obtained as Eq. (6).

$$\begin{aligned} f(x) &= \operatorname{sgn}\left[\sum_{i=1}^{m} y_i \alpha_i \cdot \left(\Phi(x) \cdot \Phi(x') \right) + b \right] \\ &= \operatorname{sgn}\left[\sum_{i=1}^{m} y_i \alpha_i \cdot k(x, x') + b \right], \\ & y_i :\ \pm 1 \\ & \alpha_i :\ \text{Lagrange multiplier} \\ & k(x, x') :\ \text{kernel, similarity of two examples } x \text{ and } x' \end{aligned} \tag{6}$$

Where k is the kernel that evaluated on input patterns x,x'. Usually there are two kinds of kernel function, i.e., polynomial and radial basis functions (rbf), expressed as Eq. (7) and Eq. (8), respectively.

$$\text{poly:} \quad k\left(x, x'\right) = \left(\left\langle x \cdot x' \right\rangle + 1\right)^d \tag{7}$$

$$\text{rbf:} \quad k\left(x, x'\right) = \exp\left(-\frac{\left\| x - x' \right\|}{2\sigma^2}\right) \tag{8}$$

Each given face image in the face database with known labels (classes) is modeled by SVM. Several images of the same person under different poses and illuminations are used for training. Recognition is accomplished by matching a test face image with unknown labels against all trained image in the face database.

3.2 One-against-all (OAA) of SVM for multi-class recognition

Basically, the OAA strategy uses a system of N_s binary SVM, where N is the class numbers. More specifically, it involves a parallel architecture made up of N numbers of SVM, and each binary SVM solves a two-class problem defined by one class against all the others. One of the N_s SVM is trained with all of the examples in the *ith* class with positive labels, and all the examples in the other classes except *ith* with negative labels. Thus given k training data $(x_1,y_1),\ldots,(x_k,y_k)$, where $x_i \in R^d$, i=1,…,k and $y_i \in \{1,\ldots,N\}$ is the class of x_i, the *ith* SVM solves the following problem:

$$\begin{aligned}
\min_{w^i, b^i, \xi^i} \quad & \frac{1}{2}(w^i)^T w^i + C\sum_{j=1}^{k} \xi_j^i \\
& (w^i)^T \varphi(x_j) + b^i \geq 1 - \xi_j^i, \ \text{ if } y_j = i, \\
& (w^i)^T \varphi(x_j) + b^i \leq -1 + \xi_j^i, \ \text{ if } y_j \neq i, \\
& \xi_j^i \geq 0, j = 1, \ldots, k
\end{aligned} \tag{9}$$

Where the $\phi(x_j)$ is a kernel function that mapped the training data x_j to a higher dimensional space and C is the penalty parameter. When data with noise causes hard margin, there is a penalty term $C\sum_{j=1}^{k}\xi_j^i$ which can relax the hard margin and allow a possibility of mistrusting the data. It can reduce the number of training errors.
After solving Eq. (9), there are N decision functions as shown in Eq. (10)

$$\begin{gathered}
f_1 = (w^1)^T \varphi(x) + b^1, \\
\vdots \\
f_N = (w^N)^T \varphi(x) + b^N,
\end{gathered}$$

$$m(x, f_1, f_2, \ldots, f_N) = \underset{i=1,\ldots,N}{\arg\max}(f_i) \tag{10}$$

where f_i is a confidential rate of the output of *ith* SVM, x is the input data, and m is the final decision that found which class has the largest value of the decision function.

3.3 One-against-one (OAO) of SVM for multi-class recognition

In the OAO strategy, several binary SVM are constructed, but each one is constructed by training data from only two different classes. Thus, this method is sometimes called a "pairwise" [32,33] approach. For a data set with N different classes, this method constructs $C_2^N = N(N\text{-}1)/2$ models of two-class SVM. The method also is very popular among researchers in neural networks, Adaboost, decision trees, etc. For training data from the ith and jth classes, we solve the following binary classification problem:

$$\begin{aligned} \min_{w^{ij},b^{ij},\xi^{ij}} \quad & \frac{1}{2}(w^{ij})^T w^{ij} + C\sum_t \xi_t^{ij} \\ & (w^{ij})^T \varphi(x_t) + b^{ij} \geq 1 - \xi_t^{ij}, \ \text{ if } y_t = i, \\ & (w^{ij})^T \varphi(x_t) + b^{ij} \leq -1 + \xi_t^{ij}, \ \text{ if } y_t = j, \\ & \xi_t^{ij} \geq 0, \end{aligned} \tag{11}$$

The simplest decision function is a majority vote or max-win scheme. The decision function counts the votes for each class based on the output from the $N(N\text{-}1)/2$ SVM. The class with the most votes is the system output.

In the majority vote process, it is necessary to compute each discriminate function $f_{ij}(x)$ of the input data x for the $N(N\text{-}1)/2$ SVMs model. The score function $R_i(x)$ is sums of the correct votes. The final decision is taken on the basis of the "winner takes all" rule, which corresponds to the following maximization. The expression for final decision is given as Eq. (12).

$$\begin{aligned} & f_{ij}(x) = (x * w^n) + b_n, \quad n = 1,\ldots,N \\ & R_i(x) = \sum_{\substack{j=1 \\ j\neq i}}^{N} \operatorname{sgn}\left\{ f_{ij}(x) \right\} \\ & m(x, R_1, R_2, \ldots, R_N) = \underset{i=1,\ldots,N}{\arg\max} \left\{ R_i(x) \right\}. \end{aligned} \tag{12}$$

where $f_{ij}(x)$ is the output of $ijth$ SVM, x is the input data and m is the final decision that found which class has the largest voting from the decision function $f_{ij.}$

3.4 Our multi-stage classifier system for multi-class recognition

Based on coarse-to-fine strategy, the proposed novel scheme for face recognition is a consecutive multi-stage recognition system, in which each stage is devoted to remove a lot of false classes more or less during the recognition process. The flowcharts of the proposed system including the training phase and recognition phase are shown in Fig. 3(a) and (b), respectively. In the first two stages (OAA and OAO of SVM), the obtained DCT features from feature extraction process are used. In the first stage (OAA), we selected the classes whose confidential rate is greater than the threshold, $f_i > T$, which are the subset for the second stage. In the second stage (OAO), the voting strategy is employed to select the top two classes with maximum votes, because the selected top two classes can reach a recognition rate nearly up to 100%. In the third stage, the Euclidian distance of each image of the two classes is calculated, and then the lowest distance of the image of each class is determined for the last stage, i.e., stage four. "RANSAC" method is applied in the last stage, in which the epipolar geometry method with space information of the testing image is matched with the two training images, and then the classified image with the greatest matching numbers of corresponding feature points is selected as the correct one ; in other words, the training image with the most geometric similarity to the testing image is thus presented.

As mentioned above, the purpose of using DCT is for dimensionality reduction, and the selected low-frequency DCT coefficient vectors are fed into the first two stages for training and testing phase (see Fig. 3(a), (b)). There are two advantages of using the same feature vectors for the first two stages, one is to save the computation cost and the other is to simplify the framework of the recognition system. Essentially, the major difference between the DCT, PCA, and RANSAC are the methodologies of feature extraction while considering what kind of features is the best for classification. For examples, DCT is for frequency transformation, PCA is a statistical technique, and RANSAC is for space geometric information. Three methods can complement each other in our recognition system. The outputs of each stage are collected by a cascade manner as shown in Fig. 3(b), and the final output of last stage accomplished by "RANSAC" method catches a striking and attractive classification end.

In the first stage, OAA testing phase, (see Fig. 3(b)), a scheme with a system of N_s binary SVM where N is the numbers of class is used. As indicated by Eq. (10), there are N decision functions, and f is a confidential rate which is an output from binary SVM, in which $f \in R$ presents. And the formula, $m(x, f_1, f_2, ..., f_N) = \arg\max_{i=1,...,N}(f_i)$, means the most confidential rate. We select a subset part of classes A_j whose confidential rate is larger than the threshold T_j, $A_j = \{f_i \mid f_i \geq T_j\}$, where $i = 1,2,\ldots,N$; $j = 1,2,\ldots,k$; N are class numbers; k are testing data numbers, $A_j \subset \{C_1, C_2,\ldots,C_N\}$, C_i = ClassLabels. Then the subset A_j of the class numbers is delivered into second stage in order to shorten the computation time. For example, only the top 10 classes whose confidential rates excess a preset threshold are preserved if there are 40 classes in all; in other words, only the top 10 of the total 40 classes are selected for use in the second stage.

In the second stage OAO model as shown in Fig. 3(b), there are $C_2^N = N(N\text{-}1)/2$ models of "pair-wise" SVM. From Eq. (12), $m(x, R_1, R_2, \ldots, R_L) = \arg\max_{i=1,...,L}\{R_i(x)\}$ is the largest voting from the decision function f_{ij}, where L is the class numbers of subset A, which is screened from first stage. In order to reduce testing error, top two classes with the first and second high voting value are selected, and the difference between their voting values is computed accordingly. If the numerical difference falls less than or equals e, where e is a setting number, these selected two classes are delivered into the third stage for binary classification; $R_i(x) - R_j(x) \leq e$, where $R_i(x)$ and $R_j(x)$ are the classes with the first and second voting value. However, if $R_i(x) - R_j(x) > e$, the *class i* is then be decided as the only correct answer. While the voting value difference falls less than or equals e, it represents a very close difference between *classes i* and *j*, it also tells that there is definitely a need to proceed to the next stage to identify the decision.

In the third stage, the eigenvectors binary model is shaped by the training phase. As shown in the Fig. 4(a), the input image is projected into "face space" and the weights form vectors $\psi^T = [w_1\ w_2\ \ldots\ w_9]$. In our case, only one image of each class is selected by the process of finding the minimum Euclidian distance. Fig. 4(b) shows the Euclidian distance between input image and the ten training images. As indicated in Fig. 4(b), it represents two classes composed of ten training images in all, class 1 includes the first five images, which are images 1, 2, 3, 4, and 5, respectively; and class 2 includes the other five images, which are images 6, 7, 8, 9, and 10, respectively. Subsequently, the image with the minimum Euclidian distance from each class is picked out for the last stage. As a result, in our case, image 5 out of class 1 and image 9 out of class 2 are decided in stage three.

In the last stage, "RANSAC" method is used to match one testing image with two training images, trying to find which training image best matches with the testing image. It shows that the one with the maximum numbers of corresponding points fits best. The procedure of "RANSAC" is described as following.

a. Find Harris corners [34] in testing and training images: Shifting a window in any direction should give a large change in intensity as shown in Fig. 5(a). The change $E_{x,y}$ produced by a shift (x,y) is given by:

$$E_{x,y} = \sum_{u,v} w_{u,v} [I_{x+u,y+v} - I_{u,v}]^2 \tag{13}$$

   where w specifies the image window, for example a rectangular function: it is unity within a specified rectangular region, and zero elsewhere. A Gaussian functions: smooth circular window $w_{u,v} = \exp-(u^2+v^2)/2\sigma^2$.
   $I_{u,v}$: image intensity
b. Find putative matches: Among previously detected corner feature points in given image pairs, putative matches are generated by looking for match points that are maximally correlated with each other within given windows. Undoubtedly, only points that robustly correlate with each other in both directions are returned. Even though the correlation matching results in many wrong matches, which is about 10 to 50 percent, it is well enough to compute the fundamental matrix F as shown in Fig. 5(b).
c. Use RANSAC method to locate the corresponding points between the testing and training images: As shown in Fig. 6, the map $x \rightarrow l'$ between two images defined by fundamental matrix F is considered. And the most basic properties of F is $x'Fx = 0$ [35] for any pair of corresponding points $x \leftrightarrow x'$ in the given image pairs. Following steps was used by RANSAC method to consolidate fundamental matrix F estimation:
   Repeat
   i. Select random samples of 8 correspondence points.
   ii. Compute F.
   iii. Measure support (number of inliers within threshold distance of epipolar line).
   Choose the F with the largest number of inliers and obtain the corresponding points $x_i \leftrightarrow x'_i$ (as shown in Fig. 5(c)).
d. Count numbers of matched and unmatched feature points: The threshold distance between two corresponding points $x_i \leftrightarrow x'_I$ is set. Match counts if the distance between two corresponding points is smaller than that of the threshold; on the contrary, no match does. For any given image pairs, the successful match pairs should be the training images with the largest matching number as shown in Fig. 5 (d).

4. Experimental results

The multi-stage-based face recognition is evaluated on three face databases, i.e., the ORL, Yale and IIS face databases as shown in Fig. 7. The ORL face database contains abundant variability in expression, pose and facial details (as shown in Fig. 7 (a)), which is used as a baseline study. The Yale face database is used to evaluate face recognition methods under varying lighting conditions (as shown in Fig. 7(b)) and the IIS face database (as shown in

Fig. 7(c)) is evaluated on a great number of images of 100 subjects, each subject has 30 different images.

We conducted experiments to compare our cascade multi-stage classifier strategy with some other well-known single classifier, e.g., the SVM with OAA, OAO, and Eigenface. The experimental platforms are Intel Celeron 2.93GHz processor, 1GB DDRAM, Windows XP, and Matlab 7.01.

4.1 Face recognition on the ORL database

The first experiment is performed on the ORL database. There are 400 images of 40 distinct subjects. Each subject has ten different images taken at different times. Each image was digitized a 112 × 92 pixel array whose gray levels ranged between 0 and 255. One sample subject of the ORL face images is shown in Fig. 7 (a). There are variations in facial expressions such as open/closed eyes, smiling/non-smiling, and glasses/no glasses. In our experiments, five images are randomly selected as training samples, the other five images and then serve as testing images. Therefore, for 40 subjects in the database, a total of 200 images are used for training and another 200 for testing and there are no overlaps between the training and testing sets. Here, we verify our system based on the average error rate. Such procedures are repeated for four times, i.e. four runs, which result in four groups of data. For each group, we calculated the average of the error rates versus the number of feature dimensions (from 15 to 100). Fig. 8 shows the results of the average of four runs and the output of each stage from the multi-stage classifier, which are SVM with OAA, OAO, Eigenface, and final stage. As shown in Fig. 8, the error rates of the output of the final stage is lower than the other three types of single classifier, our proposed method obtains the lowest error rate. The average minimum error rate of our method is 1.37% on the 30 feature numbers, while the OAA-SVM is 10.50%, OAO-SVM is 2.87%, and Eigenface is 8.50%. If we choose the best results among the four groups of the randomly selected data, the lowest error rate of the final stage can achieve 0%.

4.2 Comparison with previous reported results on ORL

Several approaches have been conducted for face recognition using the ORL database [5,6,8,9,11,13,36-41,43]. The methods of using single classifier systems for face recognition are Eigenface [5,37,38,40], DCT-RBFNN [6], binary tree SVM [8], 2D-HMM [9], LDA [39], and NFS [43]. The methods of using multi-classifiers for ORL face recognition are fuzzy fisherface [11,41], and CF^{2}C [13]. Here, we present a comparison under similar conditions between our proposed method and the other methods described on the ORL database. Approaches are evaluated on error rate, and feature vector dimension. Comparative results of different approaches are shown in Table 1. It is hard to compare the speed of different methods performed on different computing platforms, so we ignore the training and recognition time in each different approach. It is evident as indicated in the table that the proposed approach achieves best recognition rate in comparison with the other three approaches. In other words, our approach outperforms the other three approaches in respect of recognition rate.

4.3 Face recognition on the Yale database

The second experiment was conducted on the Yale face database, which contains 165 centered face images of 15 individuals and 11 images per person with major variations, including changes in illumination conditions (center-light, left-light, right-light), glasses/ no glasses, and different facial expressions (happy, sad, winking, sleepy, and surprised). The original images are 256 grayscale levels with a resolution of 195 × 231. The training and testing set are selected randomly with five training and six testing samples per person at four times. Similarly, we verify the proposed system based on the average error rate obtained from our four experimental results and calculate the error rates versus the number of feature dimensions (from 15 to 100). Fig. 9 shows the results of the average of four runs and the output of each stage from the multi-stage recognition system, which are SVM with OAA, OAO, Eigenface, and final stage. As shown in Fig. 9, the error rates of the output of the final stage is lower than the other three types of single classifier. The average minimum error rate of our method is 0.27% in the 65 feature numbers, while the OAA-SVM is 2.50%, OAO-SVM is 0.82%, and Eigenface is 10.27%. If we choose the best results among the four groups of verified data, the lowest error rate of the final stage can even reach 0%.

4.4 Comparison with previous reported results on Yale

Several approaches have been conducted for face recognition by using Yale database [6,11,13,25,40]. The methods of using Yale database in single classifier systems for face recognition are 2D-PCA [40], DCT-RBFNN [6], and DCT-NNC [25]. The methods of using multi-classifier systems for face recognition are fuzzy fisherface [41], and CF^2C [13]. Here, the comparison of the classification performance of all the methods is provided in Table 2. Again, Table 2 clearly indicates that the proposed approach outperforms the other five approaches.

4.5 Face recognition on the IIS database

The IIS face database contains 3000 face images of 100 individuals. There are 30 images per subject, the images per subject include tens for frontal face, tens for left profile, and tens for right profile. Each image was digitized and presented by 175 × 155 pixel array whose gray levels ranged between 0 and 255 as shown in Fig. 7(c). The training and testing set are selected randomly. This split procedure has been repeated four times in each case. Six images of each subject are randomly selected for training, and the remaining 24 images are for testing. The result is shown in Fig. 10 and Table 3 gives evidence that the proposed method outperformed other classification techniques.

5. Conclusions

This paper presents a multi-stage classifier method for face recognition based on the techniques of SVM, Eigenface, and RANSAC. The proposed multi-stage method is based on a coarse-to-fine strategy, which can reduce the computation cost. The facial features are first extracted by the DCT for the first two stages, i.e., OAA-SVM and OAO-SVM. Through all our experiments, OAO-SVM obtained a higher recognition rate than the OAA-SVM, so in our research, we put the OAO after the OAA. Although the last stage (RANSAC) led to more accuracy in comparison with the other three stages, its computation cost more in the geometric fundamental matrix *F* estimation. In order to decrease the computation time, we need to reduce the classes and images to only two training images to match with testing image in the last stage. The key of this method is to consolidate OAO-SVM for the output of the top two maximum votes so that the decision of the correct class could be made later by RANSAC in the last stage. The feasibility of the proposed approach has been successfully tested on ORL, Yale, and IIS face databases, which are acquired under varying pose, illumination, expression, and a great quantity of samples. Comparative experiments on the three face databases also show that the proposed approach is superior to single classifier and multi-parallel classifier.

6. References

[1] R. Chellappa, C. L. Wilson, S.Sirohey, Human and machine recognition of faces: a survey, in Proc. IEEE, 83 (5) (1995) 705-740.

[2] A. Samal, P. A. Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey, Pattern Recognition, 25 (1992) 65-77.

[3] D. Valentin, H. Abdi, A. J. O'Toole, G. W. Cottrell, Connectionist models of face processing: a survey, Pattern Recognition, 27 (1994) 1209-1230.

[4] R. Brunelli, T. Poggio, Face recognition: Features versus templates, IEEE Trans. Pattern Anal. Mach. Intell., 15 (10) (1993) 1042-1053.

[5] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience, 3 (1991) 71-86.

[6] M. J. Er, W. Chen, S. Wu, High-Speed Face Recognition Based on Discrete Cosine Transform and RBF Neural Networks, IEEE Trans. Neural Networks, 16 (3) (2005) 679-691.

[7] C. Xiang, X. A. Fan, T. H. Lee, Face Recognition Using Recursive Fisher Linear Discriminant, IEEE Trans. on Image Processing, 15 (8) (2006) 2097-2105.

[8] G. Guo, S. Z. Li, K. L. Chan, Support vector machines for face recognition, Image and Vision Computing, 19 (2001) 631-638.

[9] H. Othman, T. Aboulnasr, A Separable Low Complexity 2D HMM with Application to Face Recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence, 25 (10) (2003) 1229-1238.

[10] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, S. Z. Li, Ensemble-Based Discriminant Learning With Boosting for Face Recognition, IEEE Trans. on Neural Networks, 17 (1) (2006) 166-178.

[11] K. C. Kwak, W. Pedrycz, Face recognition: A study in information fusion using fuzzy integral, Pattern Recognition Letters, 26 (2005) 719-733.

[12] A. N. Rajagopalan, K. S. Rao, Y. A. Kumar, Face recognition using multiple facial features, Pattern Recognition Letters, 28 (2007) 335-341.

[13] D. Zhou, X. Yang, N. Peng, Y. Wang, Improved-LDA based face recognition using both facial global and local information, Pattern Recognition Letters, 27 (2006) 536-543.

[14] Z. Q. Zhao, D. S. Huang, B. Y. Sun, Human face recognition based on multi-features using neural networks committee, Pattern Recognition Letters, 25 (2004) 1351-1358.

[15] A. Lemieux, M. Parizeau, Flexible multi-classifier architecture for face recognition systems, in 16^{th} Int. Conf. on Vision Interface, (2003).

[16] J. C. Peincipe, N. R. Euliano, W. C. Lefebvre, Neural and Adaptive Systems, John Wiley & Sons, Inc., 1999.

[17] A. V. Nefian, M. H. Hayes, Face Detection and Recognition using Hidden Markov Models, in Proc. IEEE Int. Conf. Image Processing, 1 (1998) 141-145.

[18] H. Othman, T. Aboulnasr, A Separable Low Complexity 2D HMM with Application to Face Recognition, IEEE Trans. Pattern Anal. Mach. Intell. , 25 (10) (2003) 1229-1238.

[19] M. S. Kim, D. Kim, S. Y. Lee, Face recognition using the embedded HMM with second-order block-specific observations, Pattern Recognition, 36 (2003) 2723-2735.

[20] S. Eickeler, S. Birlinghoven, Face Database Retrieval Using Pseudo 2D Hidden Markov Models, in Proc. of Fifth IEEE Int. Conf. on Automatic Face and Gesture Recognition, (2002) 58-63.

[21] C. W. Hsu, and C. J. Lin, A comparison of methods for multi-class support vector machines, IEEE Trans. Neural Network, 13 (2) (2002) 415-425.
[22] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, Inc., 1998.
[23] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, Large margin DAGs for multiclass classification, In Advances in Neural Information Processing Systems, MIT Press, 12 (2000) 547-553.
[24] G. Ratsch, T. Onoda, K. R. Muller, Soft Margins for AdaBoost, Machine Learning, 42 (2001) 287-320.
[25] X. Y. Jing, D. Zhang, A Face and Palmprint Recognition Approach Based on Discriminant DCT Feature Extraction, IEEE Trans. on Systems, Man, and Cyb., 34 (6) (2004) 2405-2415.
[26] M. A. Fischler, R. C. Bolles, Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, Communications of the ACM, 24 (6) (1981) 381-395.
[27] Ara V. Nefian and Monson H. Hayes III, Hidden Markov Models for Face Recognition, in Proc. IEEE Int. Conf. Acoustic, Speech, and Signal Processing, 5 (1998) 2721-2724.
[28] V. V. Kohir, U. B. Desai, Face recognition using a DCT-HMM approach, in Proc. 4th IEEE Int. Conf. Application of Computer Vision, (1998) 226-231.
[29] M. Bicego, U. Castellani, V. Murino, Using Hidden Markov Models and Wavelets for Face Recognition, in Proc. 12th IEEE Int. Conf. Image Analysis and Processing, (2003) 52-56.
[30] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik. Comparison of classifier methods: a case study in handwriting digit recognition. In International Conference on Pattern Recognition, IEEE Computer Society Press, (1994) 77-87.
[31] Ben-Hur Asa, David Horn, T.H. Siegelmann, Vladimir Vapnik, Support Vector Clustering, J. of Machine Learning Research, (2001) 125-137.
[32] D. Price, S. Knerr, Pairwise neural network classifiers with probabilistic outputs, Neural Information Processing Systems, 7 (1994)
[33] T. Hastie, R. Tibshirani, Classification by pairwise coupling, in Conf. on Advances in Neural Information Processing Syatems, MIT Press, Cambridge, MA, (1998) 507-513.
[34] C. Harris, M. Stephens, A Combined Corner and Edge Detector, In 4th Alvey Vision Conference, (1988) 147-151.
[35] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, second ed., Cambridge University Press, 2003.
[36] C. H. Chen, C. T. Chu, High efficiency feature extraction based on 1-D wavelet transform for real-time face recognition, WSEAS Trans. on Information Science and Application, 1 (2004) 411-417.
[37] B. Li, Y. Liu, When eigenfaces are combined with wavelets, Knowledge-Based Systems, 15 (2002) 343-347.
[38] T. Phiasai, S. Arunrungrusmi, K. Chamnongthai, Face recognition system with PCA and moment invariant method, in Proc. of the IEEE International Symposium on Circuits and Systems, 2 (2001) 165-168.
[39] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, Face recognition using LDA-based algorithms, IEEE Trans. on Neural Networks, 14 (2003) 195-200.
[40] J. Yang, D. Zhang, A. F. Frangi, J. Y. Yang, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence, 26 (2004) 131-137.
[41] K. C. Kwak, W. Pedrycz, Face recognition using a fuzzy fisferface classifier, Pattern Recognition, 38 (2005) 1717-1732.

[42] C. T. Chu, C. H. Chen, J. H Dai, Multiple Facial Features Representation for Real-Time Face Recognition, Journal of Information Science and Engineering, 22 (2006) 1601-1610.
[43] J. T. Chien, C. C. Wu, Discriminant waveletfaces and nearest feature classifiers for face recognition, IEEE Trans. on Pattern Analysis and Machine Intelligence, 24 (2002) 1644-1649.
[44] ORL face database, (http://www.uk.research.att.com/facedatabase.html).
[45] Yale face database, (http://cvc.yale.edu/projects/yalefaces/yalefaces.html).
[46] IIS face database, (http://smart.iis.sinica.edu.tw/).

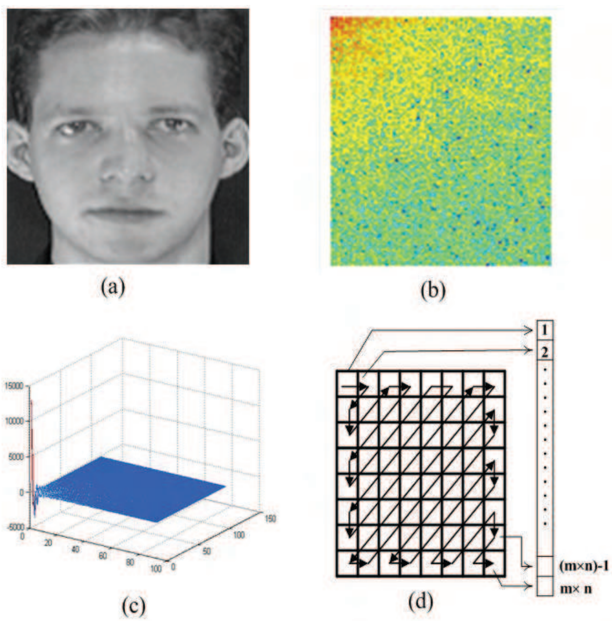

Fig. 1. Facial image and DCT transform image (a) original face image (b) 2D plot after 2D-DCT (c) 3D plot after 2D-DCT (d) Scheme of zig-zag method to extract 2D-DCT coefficients to a 1D vector

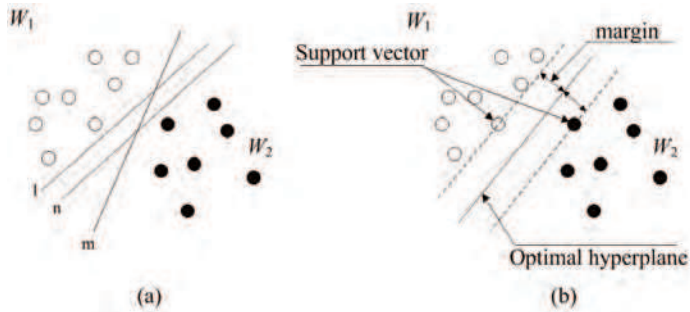

Fig. 2. Classification between two classes W_1 and W_2 using hyperplanes: (a) arbitrary hyperplanes l, m, and n. (b) the optimal separating hyperplane with the largest margin identified by the dashed line, passing the two support vectors.

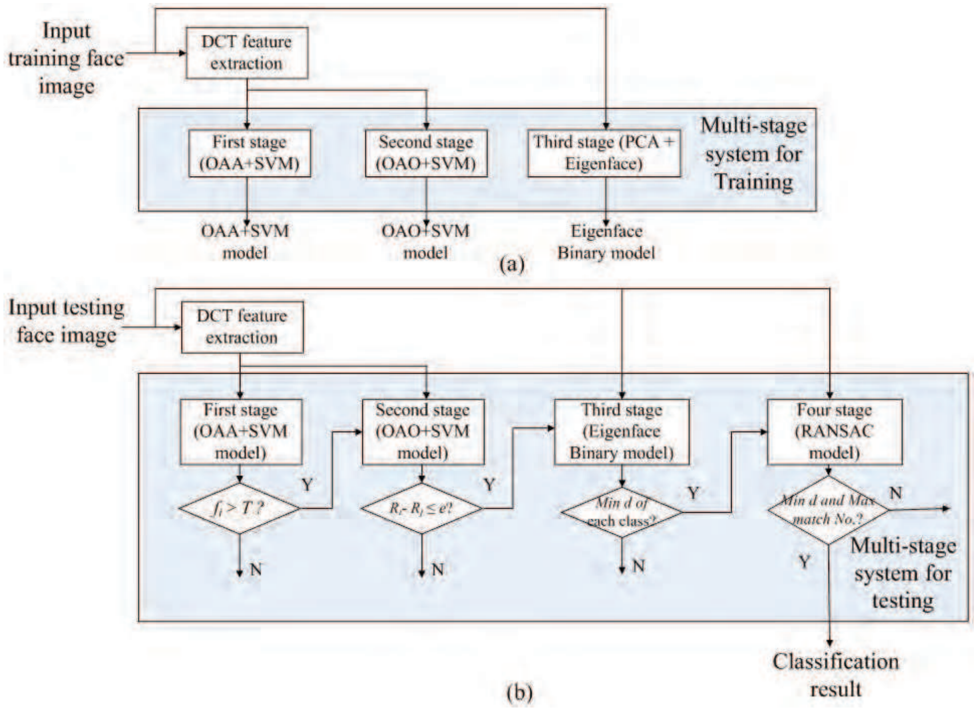

Fig. 3. Flowchart of the face recognition system. (a) training phase (b) testing phase.

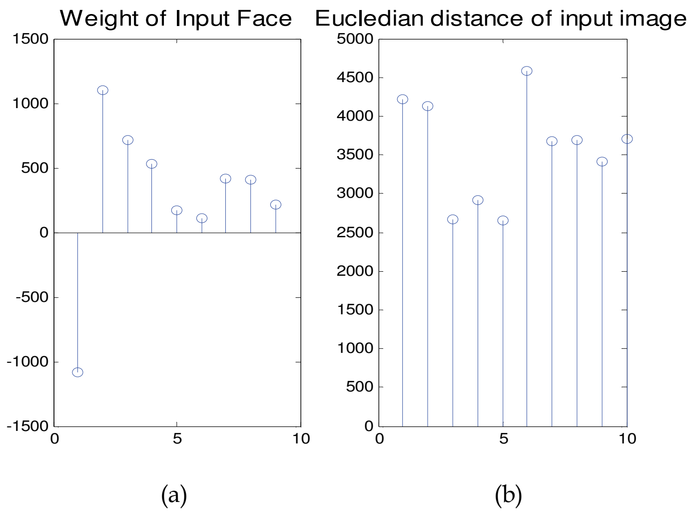

Fig. 4. (a) The weight vector $\psi^T = [w_1\ w_2\ \ldots\ w_9]$ of input image. (b) The Euclidian distance between the input image and ten training samples, respectively. The sample 1, 2, 3, 4, 5 are the same classes and sample 6, 7, 8, 9, 10 are another classes.

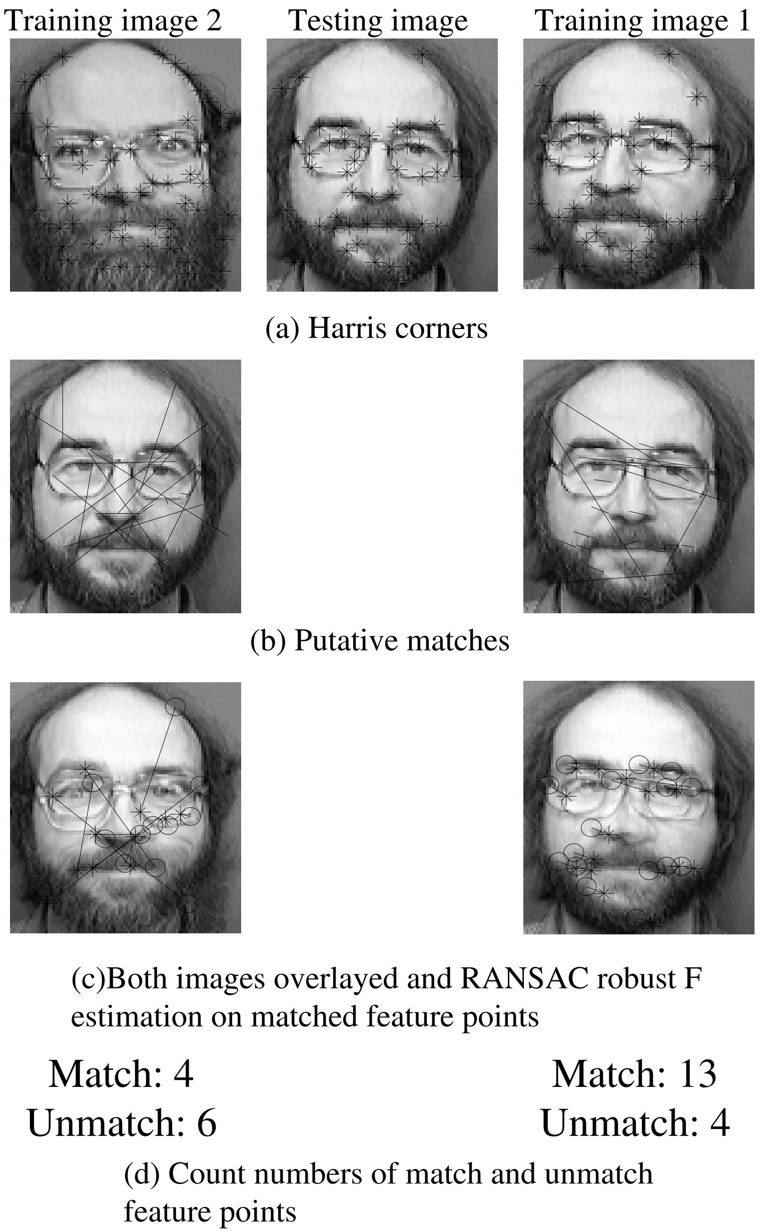

Fig. 5. Four procedures of space information using RANSAC method to find the match and unmatch feature points. (a) Find Harris corners feature points in one testing and two training images. (b) Find putative matches of testing and training images. (c) Using RANSAC method to find testing and training images of match feature points. (d) Count numbers of match and unmatch feature points.

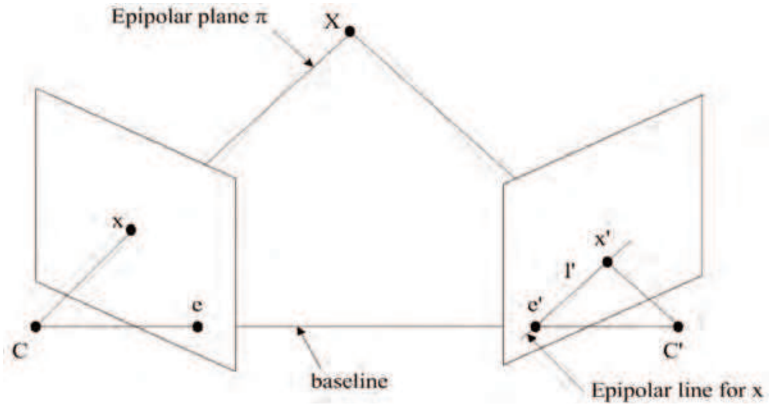

Fig. 6. Point correspondence geometry. The two cameras are indicated by their centers C and C' and image planes. The camera centers, 3-space point X, and its images x and x' lie in a common plane π. An image point x back-projects to a ray in 3-space defined by the first camera center, C, and x. This ray is imaged as a line l' in the second view. The 3-space point X which projects to x must lie on this ray, so the image of X in the second view must lie on l'.

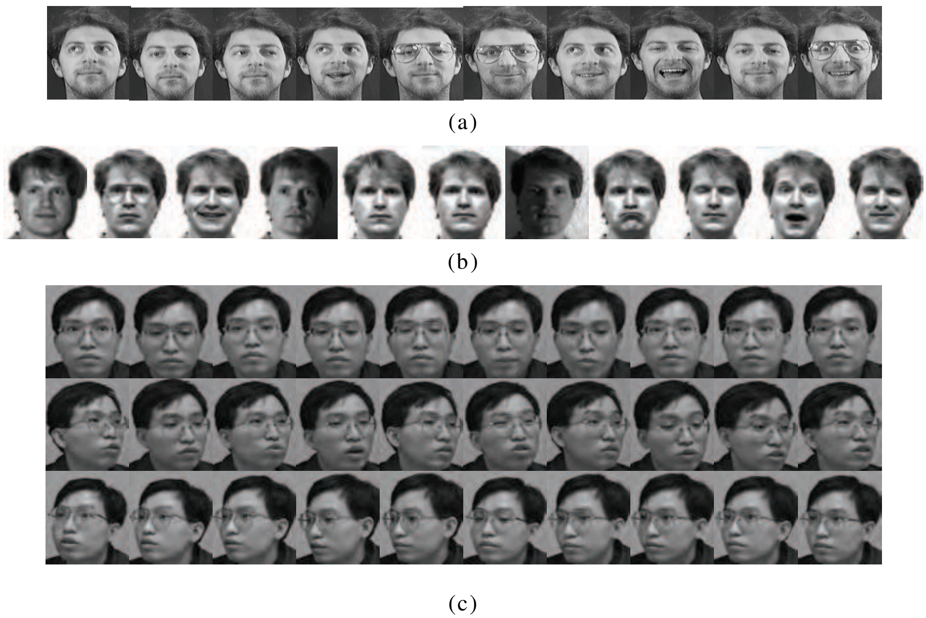

Fig. 7. Some sample images from publicly available face database used in the experiments: (a) ORL face database, (b) Yale face database, (c) IIS face database.

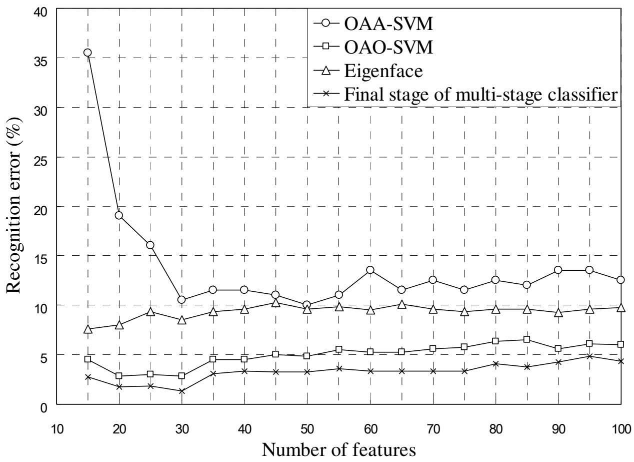

Fig. 8. Comparison of recognition error versus the number of features of the OAA-SVM, OAO-SVM, Eigenface, and final stage of the Multi-stage classifier system on the ORL face database.

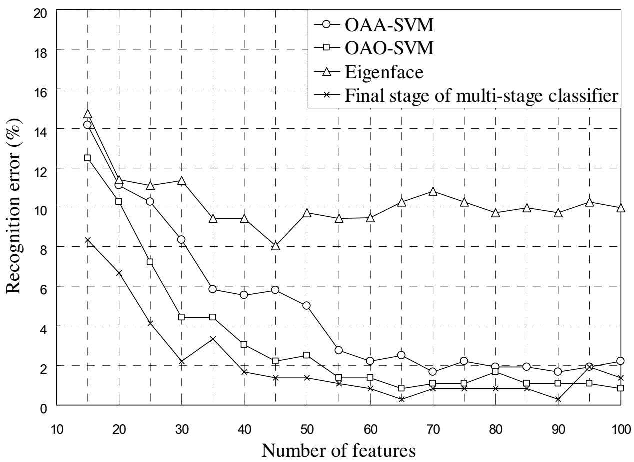

Fig. 9. Comparison of recognition error versus the number of features of the OAA-SVM, OAO-SVM, Eigenface, and final stage of the Multi-stage classifier system on the Yale face database.

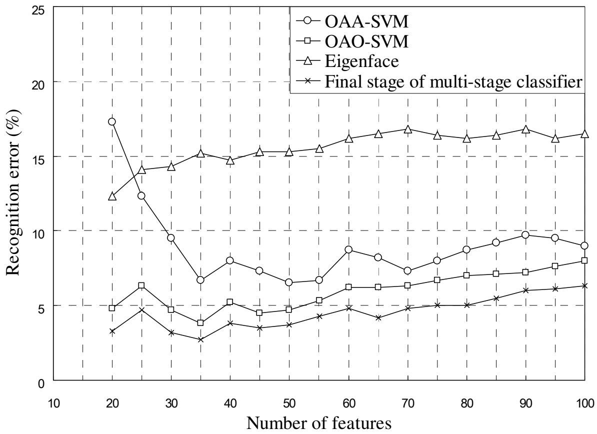

Fig. 10. Comparison of recognition error versus the number of features of the OAA-SVM, OAO-SVM, Eigenface, and final stage of the Multi-stage classifier system on the IIS face database.

Methods	Error rate (%)		Feature vector dimension
	Best	Mean	
Wavelet + Eigenface [37]	2	4	140
2D-PCA [40]	4	5	112×3
Binary tree SVM [8]	N/A	3	48
DCT-RBFNN [6]	0	2.45	30
CF^{2}C [13]	3	4	30
Fuzzy Fisherface [41]	2.5	4.5	60
Our proposed approach	0	1.375	30

Table 1. Recognition performance comparison of different approaches (ORL)

Methods	Error rate (%)		Feature vector dimension
	Best	Mean	
2D-PCA [40]	N/A	15.76	139/165
DCT-RBFNN [6]	N/A	1.8	N/A
DCT-NNC [25]	2.22	2.22	59
CF^{2}C [13]	N/A	3.1	14
Fuzzy Fisherface [41]	2.16	5.2	40
Our proposed approach	0	0.27	65

Table 2. Recognition performance comparison of different approaches (Yale)

Methods	Error rate (%)		Feature vector dimension
	Best	Mean	
Discriminant waveletface + NFS [43]	3.6	N/A	60
DWT + PNN [36]	8.83	N/A	24
Multi-feature + DWT + PNN [42]	3.08	N/A	70
Our proposed approach	1.8	2.7	35

Table 3. Recognition performance comparison of different approaches (IIS)