International Journal of Innovative Computing, Information and Control
Volume **8**, Number **7(A)**, July **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 4591–4603

A BAYESIAN FRAMEWORK FOR FACE RECOGNITION

Mohammad Reza Daliri[1,2] and Morteza Saraf[2]

[1]Faculty of Electrical Engineering
Iran University of Science and Technology
Narmak, Tehran 16846-13114, Iran
daliri@iust.ac.ir

[2]School of Cognitive Sciences
Institute for Research in Fundamental Sciences
Niavaran, Tehran 19395-5746, Iran
msaraf@ipm.ir

Received April 2011; revised November 2011

Abstract. *In this paper, a statistical face recognition scheme proposed by combining the techniques of Bayes' theorem and Parzen estimation applied on various features such as discrete wavelet transform (DWT), Discrete Cosine Transform (DCT) and Principle Component Analysis (PCA). Parzen algorithm estimates the conditional probabilities for each class and according to Bayes' theorem; the class with maximum posterior probability is selected for each test face image. The optimal Gaussian variances for each class have been found by the Genetic Algorithm (GA) optimization. The experiments on the ORL dataset demonstrate that the proposed Parzen based Bayesian classification method with enough DWT features leads, in mean recognition improvement, to 0.2% in comparison with Support Vector Machine (SVM) and 5.6% in comparison with K-Nearest Neighbour (KNN) classifier. Also applying various classifiers on DWT, DCT and PCA features, determine that with enough features of DWT, it has the best performance in compared with the others. Extra work has been performed to develop the statistical data dependence features selection in order to improve the recognition rate. This processed by searching in features space in order to minimize the reverse scattering matrix utilizing the GA. Results show that it significantly decreases the implementation complexity with selection of robust and informative features.*

Keywords: Bayesian decision rule, K-nearest neighbor, Parzen estimation, Principle component analysis, Support vector machine

1. Introduction.

Face recognition has an important task in many research areas such as security access control, identity authentication, context-based indexing. Increasing the usage of human face recognition in many real-world applications has been involved to propose many different recognition schemes over the past 30 years. Also the robustness of these techniques against the changes of illumination, pose, facial expression and background has been studied by a better representation of face images and also the decision rules [1-6].

Generally face recognition schemes are based on two approaches: constituent-based and face-based approach [7,8]. The constituent-based recognition is based on facial features relationship. However, extracting the facial features plays an important role on the accuracy of recognition [9-11]. Several algorithms, such as feature extraction based on integration projections, deformable template, Gabor wavelet features, have been proposed in order to find the facial features [9,12,13]. Face-based approach assumes a face as a whole, in which they represent the faces with the matrix of pixels intensity. The Principle Component Analysis (PCA) is a typical face-based technique, one of the most useful subspace methods which performs a covariance analysis between coefficients, and then find the projection directions corresponding to the largest data variation [14]. These directions are determined by the eigenvectors of the covariance matrix corresponding to the largest eigenvalues. The uncorrelated features of PCA coefficients in the subspace make it optimal for representation and reconstruction but the second-order dependencies are eliminated [15]. The other linear methods such as Locality Preserving Projections (LPP) [16] determine a face subspace with best detection of the essential face manifold structure, and also preserve the local information. LPP can recognize better than PCA by selecting the appropriate dimension of subspace and only when the multiple training samples are available for each face. Though face-based approach has been successfully used in face recognition, the robustness and recognition accuracy can be changed by pixel intensities variation.

It is well known that the other image representations such as wavelet-based or discrete cosine-based representation have many advantages. In frequency analysis domain, an image can be represented as a weighted combination of basis functions. The detailed information like edges determines in high frequency and the coarse information is in low frequency domain. Discrete Wavelet Transform (DWT) has been successfully used as a dimension reduction technique and/or as a feature extraction approach by decomposing an image into frequency sub-bands at different scales. So designing an appropriate DWT and using low frequency approximation sub-band can result in higher feature space robustness with regard to lighting variation [17,18]. Also discrete cosine transform (DCT) coefficients have powerful ability on data de-correlation for the purpose of pattern recognition, image processing including face recognition, compression analysis, communication and etc. After analyzing the DCT coefficients of an image, feature space reduction can be applied with conventional methods such as zigzag or zonal masking, but generally these approaches are not appropriate for all experiments. Hybrid feature extractors proposed such as combining the DCT transformation and the PCA with both advantages, in which DCT avoids singularity by reducing data dimension and also decreases the computation complexity of PCA [19]. [20] analyzes the features discrimination power of the DCT coefficients in order to select better features and increase the recognition performance.

After feature extracting step, face recognition performs by applying proper classifiers on these features. Looking to face recognition as a classification problem, many techniques have been proposed like Neural Networks, SVMs [21] and also statistical models like hidden Markov models (HMMs) and Gaussian Mixture Models (GMMs) [22,23]. Commonly statistical approaches such as HMM and GMM methods use the features only to describe a part of the face. These local features can be extracted by assuming the face as a number of blocks and analyze them. GMM has better robustness in comparison with HMM because it neglects the effect of spatial relations and each block is assumed independently [24,25]. SVM classifier is a state-of-the-art learning machine, which maps the feature space to some new feature space where the classes are more separable, and then attempts to maximize the margin between the separating boundaries and support vectors. The other statistical method, K-Nearest Neighbor (KNN), classifies the dataset according to posterior probabilities of each class. It estimates the conditional probability with variable resolutions in each region of training set according to its point's density.

In this paper, we proposed a novel approach based on hybrid Parzen estimation and Bayesian decision rule, which is optimized by the Genetic Algorithm (GA). For designing practical face recognition system where our prior knowledge about the face spaces is few, we proposed Parzen algorithm in order to estimate the conditional probabilities of face classes over all feature vectors. However, the main bottleneck of this process is to select the Gaussian variances for each class. Instead of selecting these values with trial and error, GA algorithm has been utilized in order to increase the reliability of the estimation for different face datasets. Then the classification can be performed according to the Bayesian decision rule. With increasing the number of feature vectors, the complexity of the algorithm significantly increases and it is not suitable for online face recognition applications. So a new statistical data dependence approach is performed in order to decrease the implementation complexity. The GA optimization is performed for finding better separable features according to scatter matrices ratio. This adaptive method is suitable for implementation of online applications with assumption of less informative features.

The rest of this work is organized as follows. In Section 2, feature representation and backgrounds on proposed data dependence feature extraction method are briefly reviewed. Classification methods including our proposed method are introduced in Section 3. Section 4 provides experimental results and discussion and finally conclusions are given in Section 5.

2. Feature Representation.

In this section, various feature extraction methods are considered. The goal of feature extraction is to find the robust and informative features for the next classification step. Feature extraction based on frequency analysis has involved by applying the transformation to the entire image and transforming them to the frequency space, then selecting some coefficients to construct the feature vectors. The most common frequency transformations are DWT and DCT. However, the subspace methods such as PCA can help us to find more informative uncorrelated features from image intensities. At the end, some preliminarily backgrounds on the proposed features extraction method are discussed.

2.1. Wavelet.

Space-frequency localization and multi-resolutions features of wavelet transform make it a popular tool in signal processing and image processing. The discrete WT can be achieved by passing the image through a series of filter bank stages. Each stage is created by first filtering in the horizontal dimension, and then by filtering in the vertical direction. The output of each filter then is down-sampled by a factor of 2 in order to reduce the coefficients number. With the low- and high-pass assumption of the filters created by any types of wavelet functions, DWT decompose the image into four frequency sub-bands LL, LH, HL, HH. The coarse information can be expressed by low frequency sub-bands and the detailed features appeared in high frequency sub-bands. In order to reduce the feature dimensions, we can apply N times the DWT on image and select the dc coefficients to have an N-level decomposition [17, 18].

2.2. DCT.

DCT, which is an interesting class of feature extraction approaches, attempts to decorrelate the image data. For the DCT analysis on an $M \times N$ image, the coefficients are obtained using the following formula [20]:

$$F(u,v) = \frac{1}{\sqrt{MN}}\alpha(u)\cdot\alpha(u)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\times\cos\left(\frac{(2x+1)u\pi}{2M}\right)\times\cos\left(\frac{(2y+1)v\pi}{2N}\right)$$
$$u=0,1,\ldots,M-1, \quad v=0,1,\ldots,N-1 \tag{1}$$

where

$$\alpha(\omega) = \begin{cases} \dfrac{1}{\sqrt{2}} & \omega = 0 \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

So, the DCT coefficients matrix $F(u,v)$ calculates by the terms of image intensity matrix $f(x,y)$. Then, the features vector can be selected by common approaches such as zigzag or zonal masking methods. With this transform, the frequency space is divided into three regions; the low frequencies describe the coarse information, and also are correlated with the illumination conditions [20]. The detail information appears in high frequency bands and the middle frequency bands contain useful information of the image structure and these frequencies are useful for recognition applications.

2.3. **PCA.** Principal component analysis has been extensively used as a feature extraction and/or reduction technique which is also known as the Karhunen-Loéve transform [14]. In order to reduce the feature dimensions by PCA, first the d-dimensional mean vector and covariance matrix are computed for the entire of dataset. Next the eigenvectors and eigenvalues are computed and sorted. The eigenvectors correspond to larger eigenvalues are selected to make the matrix A, and then the mean zero new features are generated by multiplying of the matrix A and mean zero feature vectors [3].

2.4. Backgrounds on the proposed data dependence feature extraction method.

2.4.1. *Scatter matrix.* The scatter matrix specifies the dispersion of learning samples around their mean. The within-class-scatter S_w and between-class-scatter S_b can be measured by knowing the covariance matrix and the prior probability of each class. $trace\{S_w\}$ is a measure of the variance of the features in all classes and $trace\{S_b\}$ is a measure of the mean in each class from the global mean [26]. With these definitions, the mixture scatter matrix S_m can be defined by a measure of variance of all training samples from the global mean. To understand the effectiveness of the classes in terms of their means, the ratio of $trace\{S_m\}$ to $trace\{S_w\}$ should be as large as possible [26]. The better features for classification can be selected by optimizing this ratio through the optimization algorithms such as GA.

2.4.2. *Genetic algorithm.* Genetic algorithm, which inspired from the principles of natural evolution, has been popular in last four decades in order to optimize problems in many engineering and science applications. It is a stochastic search technique with the advantages of not requiring objective function continuity or differentiability.

GA initialization is started with the chromosomes representation and fitness function definition, then based on the concept of natural selection inherent in natural genetics, the algorithm is performed the parents selection, crossover and mutation operations. In detail, the fitness function measures the closeness to the global minima solution. The best chromosomes are selected in each step and then crossover over them is performed by the statistical fitness-based methods such as probabilistic weighting sum, to generate new offsprings. Many parent selection algorithms have been proposed such as pairing from top to bottom, random pairing, roulette wheel selection, ranking selection, and tournament selection. At last, the mutation that is a stochastic operation help to search extensively and also run away from local minima, for example, adding some random gens of chromosomes by Gaussian distribution noise. The robustness of these techniques is followed by some correct parameter initialization like population, crossover and mutation parameters in order to find the better minima.

3. Classification Methods.

3.1. Bayesian decision theory and Parzen estimation.

Our purpose is to estimate the conditional probability densities without much prior knowledge of these densities by Parzen estimation. The procedure of estimation based on the labeled training set T_s, begins by splitting the training set into K subsets T_k, including N_k samples in the class of ω_k. In order to estimate the conditional density $p(z|\omega_k)$ for any arbitrary point z, a common way is to partition the measurement space into a finite number of bins R_i and count the number of points in each bin. So the probability density is proportional to that count and is estimated as [26]:

$$\hat{p}(z|\omega_k) = \frac{N_{k,i}}{volume(R_i) \times N_k} \text{ with } z \in R_i \tag{3}$$

where $N_{k,i}$ is the number of samples with class ω_k that fall within the ith bin. This approach works fine if the number of samples within each bin is sufficiently large. Hence, efficiency of estimation decreases with increasing the feature space dimension or having small training set. Refining the histogram by first considering only one sample from the training set like $z_j \in T_k$, we are certain that the density at this position is nonzero, $p(z_j|\omega_k) \neq 0$. Since $p(z|\omega_k)$ is continuous over the entire feature space, the density is nonzero in a small neighbourhood of z_j, but by moving away from z_j, our certainty become less. Parzen estimation utilized this idea in order to represent the knowledge of the observation of z_j by a function positioned at z_j and with an influence restricted to a small vicinity of z_j [23]. The final estimation yields with summing the function or kernel of all vectors in the training set. Let $\rho(z, z_j)$ be a Euclidian distance measure in the feature space. As the contribution of single observation, kernel $h(\rho(z, z_j))$ defines such that has its maximum at $z = z_j$ also must be monotonically decreasing as $\rho(.,.)$ increase and must be normalized to one. The final parzen estimate yields by summing of all observations:

$$\hat{p}(z|\omega_k) = \frac{1}{N_k} \sum\nolimits_{z_j \in T_k} h(\rho(z, z_j)) \tag{4}$$

In N dimensional case by using Gaussian kernel we have:

$$\rho(z, z_j) = \sqrt{(z - z_j)^T C^{-1} (z - z_j)} \tag{5}$$

$$h(\rho) = \frac{1}{\sigma_h^N \sqrt{(2\pi)^N |C|}} \exp\left(-\frac{\rho^2}{2\sigma_h^N}\right) \tag{6}$$

However, the actual choice of C is less important if the training set is very large.

After estimating the conditional probabilities of each class, Bayesian decision theory which is a fundamental statistical approach, plays a crucial role in face recognition problem. In multiclass classification, considering the effect of the observed points in the feature space as a conditional probability, bayes' theorem evaluate the uncertainty of each class ω_k after the observation in the form of posterior probability $p(\omega_k|z)$, which takes the form [26]:

$$p(\omega_k|z) = \frac{p(z|\omega_k)p(\omega_k)}{p(z)} \tag{7}$$

which the conditional probability $p(z|\omega_k)$ expresses how probable the observed data set is in the different range of vector ω_k. The dominator plays the role of normalization and can be expressed in terms of prior distribution and conditional probabilities as given by: $p(z) = \sum p(z|\omega_k)p(\omega_k)$, so the class with the highest posterior probability can be selected as a winner class.

So the proposed classification procedure can be obtained by these three steps:

1. Determine the σ_h such that the log-likelihood of the training set can be maximized:
$$\sum\nolimits_{k=1}^{K} \sum\nolimits_{j=1}^{N_k} \ln\left(\hat{p}(z_{k,j}|\omega_k)\right) \tag{8}$$
2. Estimate the conditional probabilities of each class by Parzen algorithm.
3. Assign the sample's class with maximum posterior probability according to the Bayes' theorem.

3.2. **KNN.** k-nearest neighbor classification first estimates the conditional probability with variable resolutions in each regions of the training set according to its points density. So it causes balanced between resolution and variance. KNN assumes a hypersphere around feature vector such that it contains exactly k samples from the training set. Next, it counts the number of samples found with each class from the k-nearest neighbors of that feature vector. Let r_1^k be the number of kth class point selected from the hypersphere and r_2^k be the entire number of kth class points in training set. Conditional probability estimation of kth class is proportional to the ratio of r_1^k to r_2^k [26]. With the assumption of prior probability and conditional probability, the Bayes' theorem classifies each feature vector such that the highest posterior probability can be achieved.

3.3. **SVM.** Support vector machine becomes popular for solving problems in classification and regression. SVMs solve the two-class classification problem as the linear models of the form $y(X) = w^T\phi(X) + b$, where $\phi(X)$ map the nonlinear separable feature space into the new linear separable feature space. Finding the model's parameters by perceptron algorithm, guarantees to find a solution by using iterative procedure, but it involves more parameter initialization sensitivity and generalization error. In order to improve the generalization, SVM is based on the smallest distance between the decision boundary and any of the training samples which is also named margin. SVM chooses the decision boundary, which the margin is maximized. Additional information on how SVMs use the Lagrange multipliers and quadratic programming can be found in [27,28]. Various methods have therefore been proposed for the problems of multiclass with combining multiple two-class SVMs. Let K be the number of classes. The one-versus-the-rest approach constructs K SVMs which the train kth model chooses the kth class as the positive examples and the remaining $(K - 1)$ classes as the negative examples, but not guarantee the balance of training set. Another approach constructs all different combinations of two-class SVMs on all possible pairs of classes and then selects the class with highest number of votes [26]. This approach, which is called one-versus-one, requires more training time in compared to the previous approach.

4. **Experimental Results and Discussion.**

In this section, we evaluate our proposed approaches compared with different face recognition schemes from feature extraction and selection to face classification algorithms. Our experiments have carried out on ORL database (http://www.cam-orl.co.uk). Figure 1 shows 4 classes of ORL database. Within ORL database (40 face classes, each consist of 10 different pose of 92 × 112 pixels images with 256 gray levels), we make five random sets of train and test sets each containing 5 images from each classes. The road map of the face recognition procedure is first extracting the feature vectors, then selecting the informative features and finally developing and testing the classifiers with train and test sets. All the experiments have been programmed and performed in the MATLAB 7.10.

The first step in face recognition schemes is to find a set of features or face descriptors which describe the whole information about the face image. The wavelet, DCT and PCA extractors are the most conventional approaches through their robustness to the intensities variations. The wavelet coefficients can be obtained by HAAR wavelet with 4-level

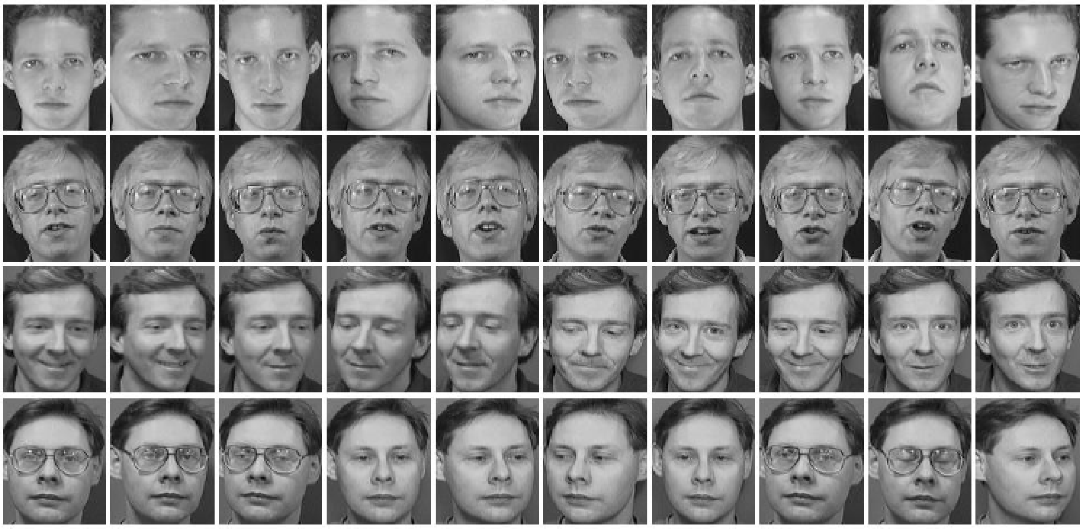

Figure 1. Sample classes of ORL dataset

decomposition, and then the features vectors are selected from the lower sub-band coefficients. DCT feature vectors can be obtained by first applying DCT on the images, and then constructing the feature vectors from DC coefficients which the method sometimes called zonal masking. Considering PCA procedure, we can find the uncorrelated features and then select the first elements as the feature vectors.

In order to find out the index of each feature vector (classification step), we proposed novel Bayesian classifier based on Parzen estimation which is optimized by GA. In the learning procedure, the conditional probabilities of the features space were estimated utilizing Parzen estimation algorithm. Each indexed conditional probability expresses how probable the observed feature vector is in the different range of that vector. We select the $\rho(z, z_j)$ as an Euclidian distance measure in the feature space and then the contribution of single observation is computed by Equation (6). The final Parzen estimation can yield by summing of all learning observations (Equation (4)). The important part of Parzen estimating is to evaluate the Gaussian density variances. If the σ_h values are selected too small, the sensitivity of the algorithm is too high and it considers the high frequency information from the data. Conversely, the large σ_h values cause to estimate the low frequency information. The optimal σ_h values can be evaluated by maximizing the log likelihood of the training dataset through Equation (8). So the task is to find the Gaussian variance for each class in order to maximize the log-likelihood function. We propose GA optimization to solve the reverse log-likelihood minimization. The chromosomes contain 40 genes each defines the indexed class Gaussian variance value. Then the GA runs in order to minimize the cost function in terms of these variances by the procedures of roulette wheel selection, crossover, and mutation with appropriate parameters, which are achieved by trial and error. Mutation helps to extend our space search and runs away from local minima by adding Gaussian distribution with random selected genes. After optimization by learning datasets, we calculate the conditional probabilities for test datasets by Parzen estimation. Then according to bayes' theorem, each sample's class has been assigned with maximum posterior probability. Experimental results determine that with fewer number of features, DCT recognizes better than PCA and DWT features; but with increasing the number of features (more than 36), DWT has the best performance (see Figure 2). Our proposed classification method with only 70 DWT features reaches to 98.7% recognition rate.

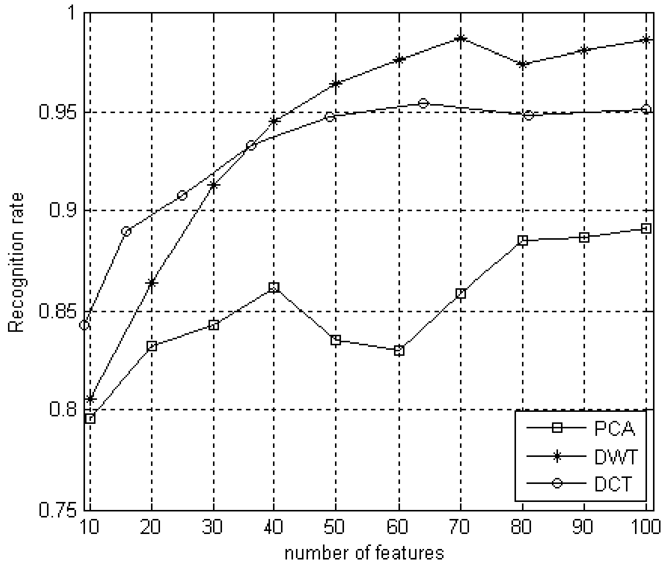

FIGURE 2. Mean of face recognition rate based on Bayesian-Parzen classification with various feature extraction methods

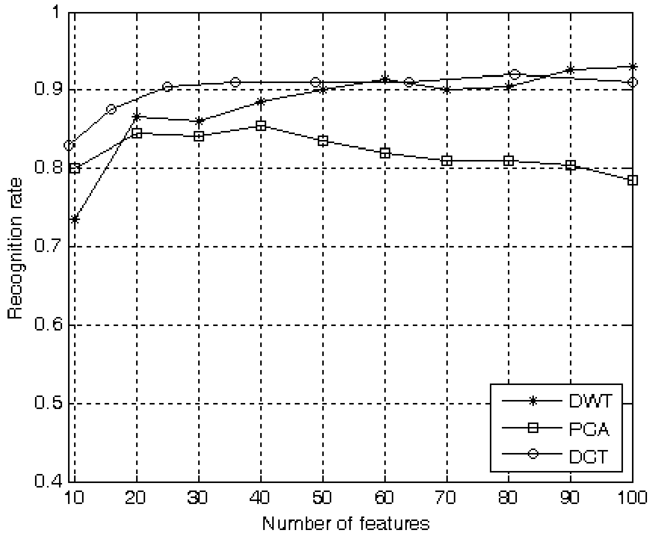

FIGURE 3. Mean of recognition rate based on KNN classification in terms of number of various features

Another way to evaluate the conditional probabilities with variable resolutions in each region of training set is KNN method. By selecting desirable k value, the conditional probability of the kth class is estimated with ratio of r_1^k to r_2^k. Then the posterior probabilities can be estimated with prior knowledge of the dataset (prior probabilities). Figure 3 presents the recognition rate in terms of the number of features with various features using KNN classifier. The low number of neighbours around two significantly increases the recognition rate. In practice, it has been suggested to make the neighbour's numbers proportional to $\sqrt{N_k}$. The results demonstrate that DWT and DCT recognize the test faces better than PCA features.

Considering SVM classifier implementation, the default MATLAB SVM package classifies two classes and maps the dataset into space kernel using linear kernel or dot product. For solving the 40-class classification, the one-versus-one method has been developed by

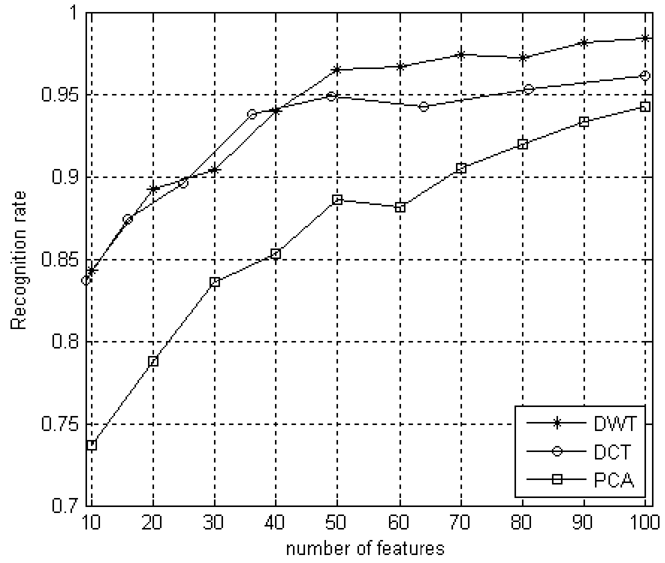

Figure 4. Mean of recognition rate based on SVM with various feature extraction methods

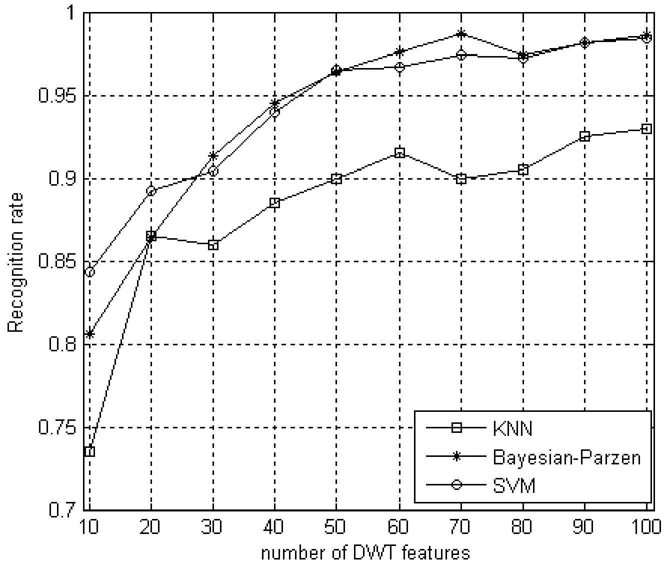

Figure 5. Mean of recognition rate based on various classification methods with DWT features

constructing all different combinations of two-class SVMs on all possible pairs of classes, and then selects the class with highest number of votes. This method requires more training time comparing with one-versus-the-rest but it guarantees the balance of training set. Figure 4 shows the SVM performance with various features extraction. SVM with DWT has the best performance of 98.4%.

The results demonstrate that enough DWT features have better performance in comparison with DCT and PCA features. Figure 5 compares the performance of classifiers in terms of number of DWT features. Considering the average recognition rate utilizing DWT, the proposed parzen based Bayesian approach leads to 0.2% improvement in comparison with SVM and 5.6% improvement in comparison with KNN classifier. Table 1 determines the Mean recognition rates with its CPU time with assumption of 100

Table 1. Evaluation of mean performances with various classification approaches and their complexities

Methods	Best Recognition Rate	STD	CPU time (s)
SVM	98.4	0.2	57.2274
KNN	93.0	0.34	5.6791
Baysian-parzen	98.6	0.74	28.5850

DWT features. It demonstrates that the complexity of proposed approach with higher recognition rate is less than SVM.

Considering the generality and dependency between recognition and datasets, we tried to find better and data dependence features for improving the recognition rates by selecting some wavelet and DCT features with maximum scattering ratios. Conventional feature selection approaches are very popular in face recognition experiments but there is no theoretically background in which these are suitable for recognition applications. In order to understand the effectiveness of the classes in terms of their means and to find better informative features for face images, the reverse of the ratio of $trace\{S_m\}$ to $trace\{S_w\}$ was used for the procedure of optimization. The search space of the optimization problem was all DWT and/or DCT coefficients. Then through the optimization, the features which were more informative were selected.

We proposed GA to solve this optimization problem. The gens of chromosome were the numbers between 1 to 10 which denotes that how many features should be selected or omitted; also the number of gens in each chromosome can be adaptive or static. Even gens denoted these next features should be selected and odd gens denote the omitted features. By this procedure, we can find better separable features according to scattering matrices ratio. The mating, mutation and populations parameters affect on the performance and should be evaluated by try and error. Also in order to extend our searching space, the chromosomes have been made with variable number of genes from 8 to 24. Table 2 summarizes the Bayesian-Parzen recognition results and the number of selected DWT and DCT features by GA-based optimization of scattering ratio. Also after the procedure, we performed the t-test on the entire dataset pair features with the significance probability of 0.05 to understand really how the features selected are separable corresponding to datasets. The null hypothesis was kept because we failed to reject it. The results determine that performance increases with fewer numbers of features with less complexity, and so this data dependence approach is useful for offline recognitions implementations with less complexity such as mobile recognition systems.

In this paper, we have proposed a novel Bayesian framework in order to solve the problem of face recognition. In order to show the efficiency of the proposed algorithm, we compared the results of the proposed method with the three methods based on neural network and statistical approaches proposed in the literature. In [29], a hybrid neural network model has been proposed for the problem of face recognition, which combines

Table 2. Performance and complexity measurements of applied feature extraction technique base on scattering ratio on face recognition rates

Features num	**DCT**		**Features num**	**DWT**	
	Mean	**Std**		**Mean**	**Std**
7.36 ± 1.1	83.5	0.15	8.46 ± 0.6	80.7	0.26
16.7 ± 0.9	90.3	0.06	20.14 ± 0.8	88.3	0.18
27.3 ± 1.3	94.3	0.06	30.1 ± 1.2	93.3	0.24

the local image sampling, a self-organizing map (SOM), and a convolutional network (CN). The role of SOM is the reduction of the dimension such that it maps a high dimensional sub-image space to a lower dimensional discrete space. They also used PCA which also known as Eigenfaces for the feature extraction phase.

In [20], the Discrimination Power Analysis (DPA) was used for the face recognition problem which is a statistical analysis based on the DCT coefficients properties and the discrimination concept. It searches for the coefficients, which have more power to discriminate different classes better than others. The best recognition rates of these methods are shown in Table 3. Considering the recognition rates, the simulations results of the various methods on ORL dataset confirm the success of the proposed method. The neural networks are so sensitive with parameter initialization and for larger number of classes; the learning time of the algorithm is too high. DPA and Eigenfaces methods have less complexity in compared to neural networks but with less performance. In our proposed approach, Gaussian variance and feature selection has an important role for increasing the performance but it is a time consuming task for the training phase because of solving the optimization problem. The procedure after training can be very fast over the test data.

Table 3. The best recognition rates of various classification approaches

Methods	**Recognition rates**
SOM + CN [33]	96.2%
Eigenfaces [33]	89.5%
DPA [22]	95.2%
Baysian-parzen	**98.7%**

In this study, a Bayesian framework for face recognition has been proposed. Bayes classifier utilizes the Parzen estimation for evaluating the posterior probabilities of each class. The advantage of Parzen estimation is that it estimates the probabilities of each face class for all ranges of the feature vector. However, an important step of this estimation is how to evaluate the Gaussian variances. As it has been mentioned, the performance of the face recognition algorithm is so sensitive to this value. The best procedure is to optimize the reverse log likelihood of the training set in terms of these Gaussian variances. The experimental results demonstrate that our proposed method significantly increases the recognition rates with respect to all face descriptors. The proposed method and the SVM classifier are so accurate but they need time for learning the face spaces and they are suitable for off-line learning applications. However, for on-line learning application, KNN is so fast but with less accuracy. Also we found that DWT features are better for accurate and robust face recognition application in compared to DCT and PCA features. In the procedure of feature selection, conventional feature selection approaches are very popular in face recognition experiments but there is no theoretically background in which these are suitable for recognitions. We developed a data dependence feature selection algorithm based on scattering matrix in order to find informative features. This method needs time for learning, but for off-line applications it finds less number of features so it decreases the implementation complexity with higher accuracy. In order to make the optimization problem for both classification and feature selection steps, our future work is to apply the PSO, SA and other evolutionary algorithms in order to find faster the local minima of the cost function.

5. Conclusions.

In this study, we implemented different face recognition schemes which have three features: 1) representation of face images by DWT, DCT and PCA features, 2) recognition by statistical and classification algorithms including the parzen estimation and Bayes classification procedure (our proposed approach), k-nearest neighbor rule, support vector machines, 3) optimizing scattering ratio utilizing genetic algorithm. The experiments on ORL database demonstrate that parzen based Bayesian classification has better recognition performance in comparison with SVM and KNN classifiers. Also comparison of various feature extraction methods on different classifiers show that enough DWT features significantly guarantee classification reliability in comparison with other features. Comparing the recognition rate of classifiers, show that the proposed approach lead, in mean improvement, to 0.2% in comparison with SVM and 5.6% in comparison with KNN. Considering feature selection approaches, the data dependence approach based on scattering ratio optimization has been developed by GA in order to improve our performance with less complexity.

Acknowledgment. The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] B. L. Zhang, H. Zhang and S. S. Ge, Face recognition by applying wavelet subband representation and kernel associative memory, *IEEE Trans. Neural Netw.*, vol.15, no.1, pp.166-177, 2004.

[2] J. Lu, K. N. Plataniotis and A. N. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, *IEEE Trans. Neural Netw.*, vol.14, no.1, pp.117-126, 2003.

[3] B. K. Gunturk, A. U. Batur and Y. Altunbasak, Eigenface-domain super-resolution for face recognition, *IEEE Trans. Image Process.*, vol.12, no.5, pp.597-606, 2003.

[4] Q. Liu, X. Tang, H. Lu and S. Ma, Face recognition using kernel scatter-difference-based discriminant analysis, *IEEE Trans. Neural Netw.*, vol.17, no.4, pp.1081-1085, 2006.

[5] W. Zheng, X. Zhou, C. Zou and L. Zhao, Facial expression recognition using kernel canonical correlation analysis (KCCA), *IEEE Trans. Neural Netw.*, vol.17, no.1, pp.233-238, 2006.

[6] X. Tan, S. Chen, Z. H. Zhou and F. Zhang, Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble, *IEEE Trans. Neural Netw.*, vol.16, no.4, pp.875-886, 2005.

[7] R. Chellappa, C. L. Wilson and S. Sirohey, Human and machines recognition of faces: A survey, *Proc. of IEEE*, vol.83, no.5, pp.705-740, 1995.

[8] A. Samal and P. A. Lyengar, Automatic recognition and analysis of human faces and facial expressions: A survey, *Pattern Recognition*, vol.25, no.1, pp.65-77, 1992.

[9] A. L. Yuille and P. W. Hallinan and D. S. Cohen, Feature extraction from faces using deformable templates, *Int. J. Comput. Vision*, vol.8, no.2, pp.99-111, 1992.

[10] S. H. Jeng, H. Y. M. Liao, C. C. Han, M. Y. Chern and Y. T. Liu, Facial feature detection using geometrical face model: An scient approach, *Pattern Recognition*, vol.31, no.3, pp.273-282, 1998.

[11] K. M. Lam and H. Yan, Location and extracting the eye in human face images, *Pattern Recognition*, vol.29, no.5, pp.771-779, 1996.

[12] H. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven and C. Malsburg, The bochum/USC face recognition system, in *Face Recognition: From Theory to Applications*, Springer, 1998.

[13] R. Brunelli and T. Poggio, Face recognition: Feature verus templates, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.15, no.10, pp.1042-1052, 1993.

[14] M. Turk and A. Pentland, Eigenfaces for recognition, *J. Cognitive Neuroscience*, vol.13, no.1, pp.71-86, 1991.

[15] L. Wiskott, J. M. Fellous, N. Krüger and C. Malsburg, Face recognition by elastic bunch graph matching, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.19, no.7, pp.775-779, 1997.

[16] X. He, S. Yan, Y. Hu, P. Niyogi and H.-J. Zhang, Face recognition using Laplacianfaces, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.27, no.3, pp.328-340, 2005.

[17] H. Sellahewa and S. Jassim, Wavelet-based face verification for constrained platforms, *Proc. of SPIE Biometric Technol. Human Identification II*, vol.5779, pp.173-183, 2005.

[18] H. Sellahewa, *Wavelet-Based Automatic Face Recognition for Constrained Devices*, Ph.D. Thesis, University of Buckingham, Buckingham, UK, 2006.

[19] W. Chen, M. J. Er and S. Wu, PCA and LDA in DCT domain, *Pattern Recognition Letters*, vol.26, pp.2474-2482, 2005.

[20] S. Dabbaghchian, M. Ghaemmaghami and A. Aghagolzadeh, Feature extraction using discrete cosine transform and discrimination power analysis with a face recognition technology, *Pattern Recognition*, vol.43, no.4, pp.1431-1440, 2010.

[21] M. Sadeghi, J. Kittler, A. Kostin and K. Messer, A comparative study of automatic face verification algorithms on the BANCA database, *Proc. of Audio-Video-Based Biometric Person Authentication*, pp.35-43, 2003.

[22] F. Samaria, *Face Recognition Using Hidden Markov Models*, Ph.D. Thesis, Trinity College, University of Cambridge, UK, 1994.

[23] S. Lucey and T. Chen, A GMM parts based face representation for improved verification through relevance adaptation, *Proc. of Int. Conf. Computer Vision Pattern Recognition*, pp.855-861, 2004.

[24] F. Cardinaux, C. Sanderson and S. Marcel, Comparison of MLP and GMM classifiers for face verification on XM2VTS, *Proc. of Audioand Video-Based Biometric Person Authentication*, pp.911-920, 2003.

[25] C. Sanderson and S. Bengio, Extrapolating single view face models for multi-view recognition, *Proc. of Int. Conf. Intelligent Sensors, Sensor Networks Information Processing*, pp.581-586, 2004.

[26] F. V. D. Heijden, R. P. W. Duin, D. De Ridder and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*, Wiley, 2004.

[27] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[28] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, 2000.

[29] S. Lawrence, C. L. Giles, A. C. Tsoi and A. D. Back, Face recognition: A convolutional neural network approach, *IEEE Trans. Neural Network*, vol.8, no.1, pp.98-113, 1997.