A New Approach in Combining Fisher's Linear Discriminant and Neural Network for Face Detection

H. FASHANDI [1], M. S. MOIN[2]
[1]Faculty of Engineering, Computer Engineering Dept.
Islamic Azad University, Science & Research Campus
Poonak, Tehran
IRAN
[2]Biometrics Research Laboratory, Information Society Group
Iran Telecommunication Research Center
P.O.Box 14155-3961, Tehran
IRAN

Abstract: This paper presents a new method for combining Fisher's Linear Discriminant (FLD) and Multi Layer Perceptron (MLP) for face detection. The input patterns are first clustered into 10 face and 10 non-face clusters using K-means algorithm. Then, the FLD coefficients are calculated to obtain the optimal projection of face and non-face images. For each 19×19 pixel window, only 19 FLD coefficients are selected and presented to a MLP classifier. The salient point of our approach, comparing with similar works, is the utilization of K-means, FLD and a simple neural network without need of preliminary transformations, such as Principle Component Analysis (PCA). The proposed method has been successfully tested on a data set completely different from the training data set, containing 970 face and 11547 non-faces. The results of experimentations exhibit an error rate of 1.34% on faces and 2.03% on non-faces, i.e. an average error rate of 1.98%, a very interesting result considering the small number of FLD coefficients and the simple structure of network, which make it an appropriate choice for real-time applications.

Key-words: - Face detection, Neural Networks, Fisher Linear Discriminant, K-means, Face recognition

1 Introduction

Face detection is the process of determining whether or not there is(are) any face(s) in a given image, and then calculating the location of each detected face, or simply labeling each region of the image into face or non-face classes. Face detection is the first step in almost any face processing system, including face identification, face verification and facial feature extraction. Due to the high degree of variations in face images, such as pose, facial expressions, lighting conditions, image orientations and imaging conditions, there are several challenges related to the problem of face detection. Obviously, the presence of color and motion in an image could facilitate the process of face detection; however, since these parameters are not common in face detection problems, to cover a wider range of applications, we focused our attention on still and gray level images.

Face detection methods can be classified in two main categories: feature-based and image-based [1][2]. Feature-based methods use explicit facial information as features and the decision is made based both on these features and the relation between them, such as geometrical distance between facial components, e.g. eyes [3]. Image based approaches use training algorithms, incorporating facial information implicitly into the system through mapping and training schemes. Examples of this category are neural networks [4], statistical approaches such as Hidden Markov Models [5] and Bayesian methods [6][7], Support Vector Machines [8], Linear subspace methods such as Principal Component Analysis (PCA) [9], Fisher Linear Discriminant and Factor Analysis [10][11][12].

Face detection by explicit modeling of facial features has been troubled by the unpredictably of face appearance and environmental conditions. By formulating the problem as a learning task to

recognize a face pattern from examples, the specific application of face knowledge is avoided. For this reason, image-based approaches are the most robust techniques for processing gray-scale static images [2].

Among image based techniques, FLD and MLP are considered as feature extraction and classification approaches, which could be used together to achieve a good performance in face detection.

Indeed FLD finds the optimal projection of face and non-face images for pattern classification, and MLP is well known as a powerful non linear classifier.

In this paper, we propose a new method for joint utilization of FLD and MLP. First, FLD features are extracted directly from raw data and then, only a small subset of features are selected and presented to a simple structure MLP with only one hidden layer.

Among previous works done on face detection, we will mention those which are particularly related to our work. In [9], distributions of face and non-face images are modeled, during the training phase, by a number of clusters (six for faces and 6 for non-faces), where each cluster is represented by a multidimensional Gaussian function with a center and a covariance matrix. In the detection phase, at each region of the input image, a feature vector is calculated (in PCA[1] space) as the distance between the local image pattern and each face/non-face cluster center, and a MLP is then used to classify these features into face or non-face classes.

In [4], first, input sub-images are pre-processed by: re-scaling them into 19×19 pixels, applying a mask for eliminating near-boundary pixels, subtracting a best-fit brightness plane from the unmasked window pixels, and finally applying histogram equalization. The intensity of each pixel is subsequently presented to a MLP with one hidden layer of 26 units, where 4 units are applied into 10×10 pixel sub regions, 16 units into 5×5 pixel sub regions and the remaining 6 units into overlapping horizontal stripes of dimension 20×5.

In [10], the training face and non-face samples are decomposed into 25 face and 25 non-face clusters using Kohonen's self-organizing map (SOM) for generating the optimal projection based on FLD. For each cluster, a Gaussian model is used to model its class conditional density function, where the parameters are estimated based on maximum likelihood. For each input pattern, the class-dependent probability is computed and the maximum likelihood decision rule is used for decision making.

In [12], training patterns of faces and non-faces are first clustered into 6 face and 6 non-face clusters using a modified K-means algorithm. Then, PCA is performed on each cluster to extract the principal components of the cluster. Two types of information are derived for each input pattern, x: (1) Its projection into PCA subspace of each cluster and (2) the normalized distance between x and its projection in each PCA subspace. Two types of FLD are performed on these 2 types of information and a vector of dimension 22 is obtained as the output of feature extraction phase. Then, two neural networks are used for classification of feature vectors in face/ non-face classes. The first network has 22 input neurons and 12 output neurons. The outputs of the first neural net are presented to the second neural net, which has only one output neuron.

As we mentioned before and will be developed more in details in the following sections, the important point of our approach, comparing with similar works, is the utilization of K-means, FLD and a simple neural networks without need of a preliminary transformation into PCA space.

The rest of the paper is organized as follows: Section 2 describes Fisher's Linear Discriminant (FLD) for feature selection. In Section 3, we present an overview of our system. Section 4, is dedicated to the experimentations and the results, and Section 5 includes concluding remarks and future works.

2 Fisher's Linear Discriminant

Selection of a representative and discriminating feature set is one of the most important phases in any pattern recognition problem. Generalization is the ability of a classifier to classify new patterns based on knowledge obtained from its previous learned patterns. Let R be the ratio of the number of learning samples to the number of free parameters of a classifier, larger R leads to better generalization performances. On the other hand, number of features, i.e. the input dimension, has a direct impact on the number of free parameters of the classifier: in general, by decreasing the number of features, the number of these parameters is also decreased and the generalization performance of the classifier is expected to be improved. Consequently, the best

[1] Principal Component Analysis

choice is a feature set, which contains the minimum number of most representative and discriminating features, leading to a maximal inter-class separability. The Fisher Linear Discriminant (FLD) method gives a projection matrix *W* that reshapes the scatter of a data set, so as to maximize inter-class separabilty, which is defined as the ratio of the "between class scatter matrix" to the "within class scatter matrix". This projection defines features that are optimally discriminating.

To calculate the matrix *W*, consider a case of *K* classes $\{C_1, C_2, \ldots, C_k\}$.We should first calculate, μ_x, the mean of all the samples as follows:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

where *N* is the total number of samples in the classes and x_i is i^{th} sample of data. The between-class scatter matrix is

$$S_b = \sum_{k=1}^{K} N_k (\mu_{xk} - \mu_x)(\mu_{xk} - \mu_x)^T \qquad (2)$$

where μ_{xk} is the mean of class *k*. The within-class scatter matrix is defined by:

$$S_W = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_{xk})(x_i - \mu_{xk})^T \qquad (3)$$

The transformation matrix *W* with the highest seperability is the one which maximizes

$$J = \frac{\det(W^T S_b W)}{\det(W^T S_W W)} \qquad (4)$$

Generalized eigenvectors of S_b and S_W, i.e. the eigenvectors of $S_b S_W^{-1}$, will maximize the value of *J*.

As seen earlier, since the number of training samples needed to obtain good generalization performances, grows with the dimension of the feature space, this dimension should be reduced in such a way that no essential information is lost. Dimensionality reduction is done by discarding eigenvectors with smaller eigenvalues. Let *x* and *y* denote an input vector and its feature vector, respectively. We define *y* as:

$$y = W_m^T x \qquad (5)$$

where W_m is a matrix containing generalized eigenvectors of $S_b S_W^{-1}$ with largest eigenvalues. Obviously, *y* is a feature vector with lower dimension than the input vector. There will be no loss in class seperabilty information, if *m=K-1*, where *K* is the number of classes [13].

The major drawback of using FLD is the singularity of S_W, caused by the problem of small sample size. Let *N* be the number of samples and *n*, the number of pixels in each image. If *n* is smaller than *N*, then S_W will be singular and generalized eigenvectors of $S_b S_W^{-1}$ could not be calculated. The rank of S_b is *K-1* or less, since it is the sum of *K* matrices of rank one or less. Similarly, the rank of S_W is at most *N-K*. For a set of *N* sample images of *n* pixels, where *N* is usually smaller than *n*, the within-scatter matrix $S_W \in R^{n \times n}$ is always singular [9]. One of the solutions to this problem is to project first the image set to a lower dimension space using PCA and then to determine *W* [9]. Another method to be used is SVD[2], which is based on a powerful technique dealing with sets of equations or matrices that are either singular, or numerically very close to singular [15]. In another approach proposed in [14], a linear transformation is applied to the original sample space *V: T(x) = S_Wx, x∈ V*. Since the rank of S_W is smaller than the dimensionality of *V* (in small sample size problems), there must exist a subspace $V_0 \subset V$ such that $V_0 = span\{\alpha_i \mid S_W\alpha_i = 0 \text{ , for } i=1,\ldots,n-r\}$, where *r* is the rank of S_W. Let $Q = [\alpha_1, \ldots, \alpha_{n-r}]$. First, all input patterns *x*'s are transformed from *V* into its corresponding subspace V_0 through the transformation QQ^t. Then, the eigenvectors corresponding to the largest eigenvalues of the between-class scatter matrix $\tilde{S}_b$ in the subspace V_0, are selected as the most discriminant vectors.

As mentioned before, FLD coefficients have been used as features in our system, which are described in the following section.

3 System Overview

Figure 1 shows an overview of the system we designed for face detection. In this section, we describe different parts of this system.

[2] Singular Value Decomposition

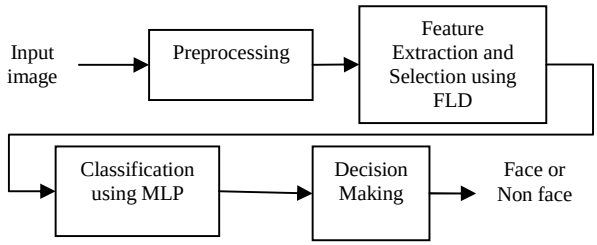


Fig. 1– Overview of face detection system

3.1 Preprocessing

The size of the scanning window for face detection is 19×19 pixels. This size permits to keep the dimensionality of the window vector space manageably small, but is still large enough to preserve distinctive visual features of face patterns. In order to detect faces with different sizes, input images are scaled consecutively during different preprocessing steps. This process locates all faces in the scanning window, as shown in Figure 2. A contrast enhancement is then done via histogram equalization.

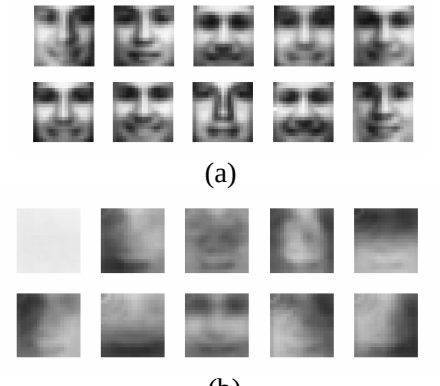
Fig. 2- Scanning window is applied to the entire image in different scales

3.2 Feature extraction and selection

As mentioned above, FLD gives the optimum features with maximum class seperability. Thus, among features that can be extracted from an image, such as DCT[3], PCA and intensity of each pixel, we have decided to select FLD coefficients.
To obtain FLD coefficients, the FLD matrix in the input space should be calculated for both face class and non-face classes. However, due to the intrinsic variations in both face and non-face classes, we first cluster them into 10 face and 10 non-face clusters. The standard K-means algorithm is then used for clustering. K-means clusters N data points, x_n, into K disjoint subsets S_j, containing N_j data points, each subject to the constraint of minimizing the sum-of-squares criterion:

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \left\| x_n - \mu_j \right\|^2 \qquad (6)$$

where μ_j is the mean, i.e. the center, of j^{th} clusters, $\|.\|$ denotes the Euclidean distance, and K is the number of clusters. Centers of face and non-face clusters are shown in Figure 3.



(a)

(b)

Fig. 3- Centers of (a) face and (b) non-face clusters

To overcome the problem of singularity of the S_W matrix, none of the approaches mentioned in section 2 have been used. Indeed, the singularity of the within-class scatter matrix is produced by the small number of training patterns N compared to the number of pixels n in the input image. All of the above mentioned techniques deal with this situation. Fortunately, we could avoid the singularity problem by collecting a considerable number of patterns for training phase, such that $N>>n$. We recall that N=6500 and n=361.

3.3 Classification

The network that we have used for classification, has 19 input neurons (based on the number of FLD coefficients), 35 neurons in hidden layer and one output neuron (Figure 4). The number of neurons in hidden layer has been determined using experimentations described in section 4.
To train the network, we have used a network training function that updates the weight and bias values according to the Levenberg-Marquardt optimization algorithm.

[3] Discrete Cosine Transform

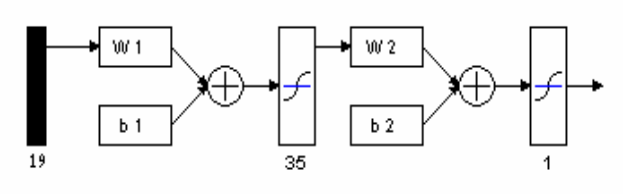


Fig. 4- Structure of the classifier.

In order to produce a network that generalizes well, it minimizes a combination of squared errors and weights. The process is called Bayesian regularization.

4 Experimental Results

To train and test our system, we have used a data set containing 8000 face and 8000 non-face samples. Each sample is a 19×19 pixel image. From this data set, first the FLD coefficients of images have been extracted, and then the first 19 coefficients have been selected. Note that for obtaining FLD matrix, we considered 10 clusters for faces and 10 clusters for non faces, giving a total of 20 clusters, and as mentioned in section 2, to avoid any loss in class seperability information, we fixed the number of FLD coefficients as $m=K-1$, where K is the number of classes(clusters). These 19 coefficients provide the input vector of the network, which is trained to have a binary output: 1 for faces and -1 for non-faces.

In order to determine the best number of neurons in the hidden layer, the network has been tested with different number of neurons during a cross-validation phase. As mentioned before, our training set contains 16000 samples of faces and non-faces. We have used 14000 of them in training set, 1000 samples for validation set and 1000 samples for test set. Fig. 5 shows the error rates of the network for faces and non-faces using different number of neurons in hidden layer for three different testing set containing 1000 samples of face and non-face each. As shown in this Figure 5 the best results were obtained using 35 neurons in hidden layer.

In another experiment, a total of 970 face and 11574 non-face images have been used to test the network performance. Some samples of this data set are shown in Figure 6. We obtained an error rate of 1.34% for faces and 2.03% for non-faces using this data set.

Figure 7 shows the results of face detection for some difficult images.

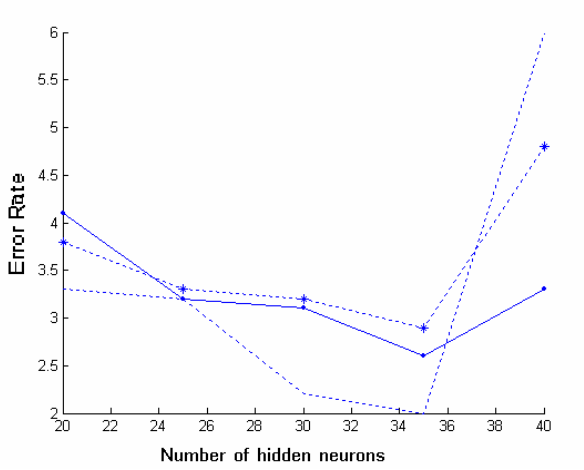


Fig. 5- Error rates of the network tested on three different data sets for different number of hidden neurons

a) Face samples

b) non-face samples

Fig. 6 - Face and non-face samples of test dataset

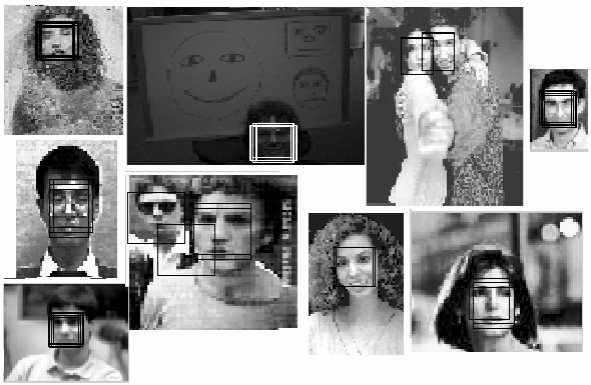


Fig. 7- Some test examples and the corresponding detected faces

4 Conclusions

In this paper, we have presented a method for face detection based on Fisher's Linear Discriminant and Multi layer Perceptron. Face detection is a very

complicated task, since: (1) there are a considerable number of non face categories, and some of them are very similar to faces and (2) there are wide variations in face images, due to the different poses, facial expressions, lighting conditions, image orientation and imaging conditions. In order to overcome this wide range of variations, a remarkable number of features are needed. However, increasing the number of features can lead us to the problem of "coarse of dimensionality" and also to a complex classifier architecture with poor generalization performance. As a solution to this problem, one should select the most discriminating features, which reduces the input space dimensionality, without loosing important information in input patterns.

In our system, Fisher's Linear Discriminant has been used to determine the optimal projection of face and non face images for pattern classification. In order to obtain a better representation of intra-class variations of both face and non faces, input images are clustered into a number of clusters using K-means clustering method, before extracting their FLD coefficients. A very small set of features are then selected amongst extracted FLD coefficients, which are used as inputs of a MLP for face, non-face classification. By decreasing the number of input features and also by applying the cross-validation technique to determine the best number of hidden layer's nodes during the MLP learning phase, we further improved the generalization performance of the face detection system. This system has been tested on a completely different set of face and non-face images, exhibiting a very low error rate, which is an indication of its good generalization performance.

In order to further improve the face detection accuracy of our system, we have planned to use "Kernel Fisher Dicriminant", which is based on non-linear directions and can be used to compensate for the shortcomings of FLD. Indeed, the distribution of training patterns is sometime such that FLD can not yield satisfactory results and a non-linear projection is needed to obtain the optimal solution.

References:

[1] M.–H. Yang, D. Kriegman and N. Ahuja, "Detecting Faces in Images: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, 2002, pp. 34-58.

[2] E. Hjelmas and B. K. Low, "Face Detection: A Survey", *Computer Vision and Image Understanding* 83, 2001, pp. 236-274.

[3] K. C. Yow and R. Cipolla, "Feature-based human face detection", *Image and Vision Computing*, 15(9), 1997, pp. 713-735.

[4] H. A. Rowley, S. Baluja and T. Kanade, "Neural Network-Based Face Detection", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, 1998, pp. 23-38.

[5] A. V. Nefian and M. H. Hayes, "Face detection and recognition using Hidden Markov Models", *International Conference on Image processing (ICIP98)*, Vol. 1, 1998, pp. 141-145.

[6] L. Meng, T. Q. Nguyen, D. A, Castanon, "An Image-based Bayesian Framework for Face Detection", *Proceedings of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2000.

[7] H. Schneiderman and T. Kanade, "A Statistical Method for 3D Object Detection Applied to Faces and Cars", IEEE *Conference on Computer Vision and Pattern Recognition*, 2000.

[8] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection", *Proceeding of IEEE conference on Computer Vision and Pattern Recognition*, 1997, pp. 130-136.

[9] K.-K. Sung, Tomaso Poggio, "Example-Based Learning for view based human face detection", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, 1998, pp. 39-51.

[10] M. H. Yang , D. Kriegman and N. Ahuja, "Face Detection Using Multimodal Density Models", *Computer Vision and Image Understanding* 84, 2001, pp. 264-284,.

[11] T. Kurita and T. Taguchi, "A Modification of Kernel-based Fisher Discriminant Analysis for Face Detection", Proc. *Of the 5th International Conf. on Automatic Face and Gesture Recognition*, 2002, pp. 300-305.

[12] Q. Gu and S. Z. Li, "Combining Feature Optimization Into Neural Network Based Face Detection", *International Conference on Pattern Recognition (ICPR'00)*, Vol. 2, 2000, pp.814-817.

[13] S. Theodoridis, K. Koutroumbas,"Pattern Recognition", *Academic Press*, 1999.

[14] L.-F. Chen, H.-Y. Mark Liao, M.-T Ko, J.-C. Lin, G.-J. Yu, 'A new LDA-based face recognition system which can solve the small sample size problem', *Pattern Recognition*, 33,2000, , pp. 1713-1726,.

[15] G. H Golub, and C.G. Van Leon, "Matrix Computations", 3rd edition, Baltimore: *John Hopkins university Press*,1996.