HOLISTIC FACE RECOGNITION BY DIMENSION REDUCTION

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

OF

THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET BAHTİYAR GÜL

IN PARTIAL FULLFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

SEPTEMBER 2003

Approval of the Graduate School of Natural and Applied Sciences

_____

Prof. Dr. Canan ÖZGEN

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Prof. Dr. Mübeccel DEMİREKLER

Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assoc. Prof. Dr. A. Aydın ALATAN

Supervisor

Examining Committee Members

Prof. Dr. Mete SEVERCAN _____

Prof. Dr. İsmet ERKMEN _____

Assoc. Prof. Dr. A. Aydın ALATAN _____

Assoc. Prof. Dr. Gözde BOZDAĞI AKAR _____

Assoc. Prof. Dr. Volkan ATALAY _____

ABSTRACT

HOLISTIC FACE RECOGNITION BY DIMENSION REDUCTION

Gül, Ahmet Bahtiyar

M.Sc., Department of Electrical and Electronics Engineering

Supervisor: Assoc. Prof. Dr. A. Aydın Alatan

September 2003, 104 pages

Face recognition is a popular research area where there are different approaches studied in the literature. In this thesis, a holistic Principal Component Analysis (PCA) based method, namely Eigenface method is studied in detail and three of the methods based on the Eigenface method are compared. These are the Bayesian PCA where Bayesian classifier is applied after dimension reduction with PCA, the Subspace Linear Discriminant Analysis (LDA) where LDA is applied after PCA and Eigenface where Nearest Mean Classifier applied after PCA. All the

three methods are implemented on the Olivetti Research Laboratory (ORL) face database, the Face Recognition Technology (FERET) database and the CNN-TURK Speakers face database. The results are compared with respect to the effects of changes in illumination, pose and aging. Simulation results show that Subspace LDA and Bayesian PCA perform slightly well with respect to PCA under changes in pose; however, even Subspace LDA and Bayesian PCA do not perform well under changes in illumination and aging although they perform better than PCA.

Keywords: Face Recognition, Principal Component Analysis (PCA), Eigenface, Linear Discriminant Analysis (LDA), Bayesian Classifier, Generalization, Discrimination, Holistic, Dimension Reduction Techniques.

ÖZ

BOYUT İNDİRGEME YOLUYLA BÜTÜNCÜ YÜZ TANIMA

Gül, Ahmet Bahtiyar

Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi: Doç. Dr. A. Aydın Alatan

Eylül 2003, 104 sayfa

Yüz tanıma, literatürde üzerine değişik araştırmalar olan popüler bir araştırma konusudur. Bu tezde, Özyüz adı verilen bütüncü bir Ana Bileşen Analizi (ABA) temelli metod detaylı olarak incelenmiş ve bu metoda dayalı 3 method karşılaştırılmıştır. Bu metodlar; ABA ile boyutu indirgenmiş resime Bayes Sınıflayıcısı uygulanması esasına dayalı Bayes ABA, ABA ile indirgenmiş resime Doğrusal Ayırtaç Analizi (DAA) uygulanması esasına dayalı Alt-Uzay DAA ve ABA ile indirgenmiş resime En-Yakın-Orta-Değer Sınıflayıcısı uygulanmasına dayalı

Özyüz'dür. Bu üç metod Olivetti Araştırma Laboratuarı (ORL) yüz veritabanı, Yüz Tanıma Teknolojisi (FERET) veritabanı ve CNN-TURK Spikerleri yüz veritabanı üzerinde uygulanmıştır. Sonuçlar aydınlatma ve yüz ifadesi değişiklikleri ve yaşlanma etkilerine göre karşılaştırılmıştır. Sonuçlar, Alt-Uzay DAA ve Bayes ABA'nın yüz ifadesi değişikliği durumunda ABA'ya göre biraz daha iyi performans gösterdiğini ortaya koymuştur. Ancak, aydınlatma değişikliği ve yaşlanma konusunda ABA'dan daha iyi sonuçlar elde edilse de Alt-Uzay DAA ve Bayes ABA'nın performansı çok iyi değildir.

Anahtar Kelimeler: Yüz Tanıma, Ana Bileşen Analizi (ABA), Özyüz, Doğrusal Ayırtaç Analizi (DAA), Bayes Sınıflayıcısı, Genelleme, Ayırma, Bütüncü, Boyut İndirgeme Teknikleri.

To my friend, Tayfun...

Dostum, Tayfun'a...

ACKNOWLEDGEMENTS

I would like to express my best wishes to my advisor Assoc. Prof. Dr. A. Aydın Alatan for his motivation, guidance, suggestions, and support anytime in the study. Also, I would like to thank to my pre-advisor Assoc. Prof. Dr. Mehmet M. Bulut for his suggestions to direct me into this field. Special thanks to my family; my father Ahmet Gül, my mother Ayşe Gül, my brothers Mesut and Mutlu Gül and my sister Fatma Fehime Gül for their great love and support on me to complete this study. I would also like to thank to Aslı Şen Dağlı for her great support and love, and her help throughout the thesis study, especially in constucting the CNN-TURK Speakers Face Database. Finally, I would like to thank to my company Aselsan and my boss Birol Erentürk for their support and sensitivity in this study.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

LIST OF ABBREVIATIONS

AI	: Artificial Intelligence
ARL	: Army Research Library
BMP	: Windows Bitmap
DARPA	: Defense Advanced Research Projects Agency
DCT	: Discrete Cosine Transform
DLA	: Dynamic Link Architecture
EBGM	: Elastic Bunch Graph Matching
FERET	: Face Recognition Technology
FLD	: Fisher Linear Discriminant Analysis
FT	: Fourier Transform
GWT	: Gabor Wavelett Transform
HMM	: Hidden Markow Model
ICA	: Independent Component Analysis
JPEG	: Joint Photographic Experts Group
LDA	: Linear Discriminant Analysis
MAP	: Maximum A Posteriori
ML	: Maximum Likelihood
NN	: Neural Network

NN-Rule : Nearest Neighbor Rule

ORL : Olivetti Research Laboratory

PCA : Principal Component Analysis

RBF : Radial Basis Function

SVM : Support Vector Machine

TIF, TIFF : Tagged Image File Format

CHAPTER 1

INTRODUCTION

Face recognition has been an active research area over the last 30 years. It has been studied by scientists from different areas of psychophysical sciences and those from different areas of computer sciences. Psychologists and neuroscientists mainly deal with the human perception part of the topic, whereas engineers studying on machine recognition of human faces deal with the computational aspects of face recognition.

Face recognition has applications mainly in the fields of biometrics, access control, law enforcement, and security and surveillance systems. Biometrics are methods to automatically verify or identify individuals using their physiological or behavioral characteristics [1]. Biometric technologies include [2]:

- Face Recognition
- Finger Print (dactylogram) Identification
- Hand Geometry Identification
- Iris Identification
- Voice Recognition
- Signature Recognition
- Retina Identification

- DNA Sequence Matching

The necessity for personal identification in the fields of private and secure systems made face recognition one of the main fields among other biometric technologies. The importance of face recognition rises from the fact that a face recognition system does not require the cooperation of the individual while the other systems need such cooperation. Thus, it should not be surprising that a face recognition system is placed in the Statue of Liberty in US.

Face recognition algorithms try to solve the problem of both verification and identification [3]. When verification is on demand, the face recognition system is given a face image and it is given a claimed identity. The system is expected to either reject or accept the claim. On the other hand, in the identification problem, the system is trained by some images of known individuals and given a test image. It decides which individual the test image belongs to.

The problem of face recognition can be stated as follows: Given still images or video of a scene, identifying one or more persons in the scene by using a stored database of faces [4]. The problem is mainly a classification problem. Training the face recognition system with images from the known individuals and classifying the newly coming test images into one of the classes is the main aspect of the face recognition systems.

The topic seems to be easy for a human, where limited memory can be a main problem; whereas the problems in machine recognition are manifold. Some of possible problems for a machine face recognition system are mainly;

1) **Facial expression change:** A smiling face, a crying face, a face with closed eyes, even a small nuance in the facial expression can affect facial recognition system significantly.
2) **Illumination change:** The direction where the individual in the image has been illuminated greatly effects face recognition success. A study on illumination effects on face recognition showed that lighting the face bottom up makes face recognition a hard task [5].
3) **Aging:** Images taken some time apart varying from 5 minutes to 5 years changes the system accuracy seriously.
4) **Rotation:** Rotation of the individual's head clockwise or counter clockwise (even if the image stays frontal with respect to the camera) affects the performance of the system.
5) **Size of the image:** A test image of size 20x20 may be hard to classify if original class of the image was 100x100.
6) **Frontal vs. Profile:** The angle in which the photo of the individual was taken with respect to the camera changes the system accuracy.

There are studies on face recognition from profile images, but this problem is out of the scope of this thesis. In this thesis, the problems stated above have been studied, except for the last item. All the images are assumed to be frontal with a maximum acceptable rotation of 10-15 degrees to the right or left.

Face recognition system is the next step of a face detection system. In this study, it is assumed that all the images used to train or test the system are face images. The problem of determination of the exact location of the face is out of the scope of this thesis, but some

preprocessing techniques had to be applied to ensure that all the faces were oriented in the same location in all the images.

1.1. Scope of the Thesis

In this thesis, the performances of three statistical face recognition techniques were studied on three different face databases. First, a holistic Principal Component Analysis (PCA) based method, namely Eigenface method [10] is studied in detail and three of the methods based on the Eigenface method are compared. These are the Bayesian PCA [24], [25], [26], [27], [28], [29], [30], [31] where Bayesian classifier is applied after dimension reduction with PCA, the Subspace Linear Discriminant Analysis (Subspace LDA) [21], [22], [23] where LDA is applied after PCA and Eigenface where Nearest Mean Classifier is applied after PCA. All the three methods are implemented on the Olivetti Research Laboratory (ORL) face database [55], the Face Recognition Technology (FERET) database [56], [57], [58] and the CNN-TURK Speakers face database. Moreover, the effect of some preprocessing techniques on the performance of these face recognition systems is investigated. The results are compared with respect to the effects of changes in illumination, pose and aging.

1.2. Thesis Outline

In Chapter 2, a brief survey on face recognition methods is given. The approach to face recognition is mainly a classification problem of pattern recognition after extraction of the features. Hence, methods on extraction of the features (both holistic and local) and common classification methods used in pattern recognition are discussed.

In Chapter 3, classification methods used in this thesis will be given in detail. Classification basics, Bayesian classifiers, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and the motivation around these topics will be covered.

In Chapter 4, the application of the methods described in Chapter 3 to face recognition shall be discussed. PCA based face recognition, LDA based face recognition, Subspace LDA (PCA + LDA) and Bayesian PCA (PCA + Bayes Classifier) methods shall be identified.

In Chapter 5, simulation setup, information about the common face databases ORL and FERET, and the results of the tests done on these databases will be given.

In Chapter 6 the subject will be concluded and the results of the simulation results shall be discussed.

CHAPTER 2

SURVEY ON FACE RECOGNITION METHODS

Face Recognition has been an interesting issue for both neuroscientists and computer engineers dealing with artificial intelligence (AI). A healthy human can detect a face easily and identify that face, whereas for a computer to recognize faces, the face area should be detected and recognition comes next. Hence, for a computer to recognize faces, the photographs should be taken in a controlled environment; a uniform background and identical poses makes the problem easy to solve. These face images are called mug shots [6]. From these mug shots, canonical face images can be manually or automatically produced by some preprocessing techniques like cropping, rotating, histogram equalization and masking.

The history of studies on human face perception and machine recognition of faces are given in this chapter.

2.1. Human Face Recognition

When building artificial face recognition systems, scientists try to understand the architecture of human face recognition system. Focusing on the methodology of human face recognition system may be useful to understand the basic system. However, the human face recognition system utilizes more than that of the machine recognition system which

is just 2-D data. The human face recognition system uses some data obtained from some or all of the senses; visual, auditory, tactile, etc. All these data is used either individually or collectively for storage and remembering of faces. In many cases, the surroundings also play an important role in human face recognition system. It is hard for a machine recognition system to handle so much data and their combinations. However, it is also hard for a human to remember many faces due to storage limitations. A key potential advantage of a machine system is its memory capacity [5], whereas for a human face recognition system the important feature is its parallel processing capacity.

The issue "which features humans use for face recognition" has been studied and it has been argued that both global and local features are used for face recognition. It is harder for humans to recognize faces which they consider as neither "attractive" nor "unattractive".

The low spatial frequency components are used to clarify the sex information of the individual whereas high frequency components are used to identify the individual. The low frequency components are used for the global description of the individual while the high frequency components are required for finer details needed in the identification process.

Both holistic and feature information are important for the human face recognition system. Studies suggest the possibility of global descriptions serving as a front end for better feature-based perception [5]. If there are dominant features present such as big ears, a small nose, etc. holistic descriptions may not be used. Also, recent studies show that an inverted face (i.e. all the intensity values are subtracted from 255 to obtain the inverse image in the grey scale) is much harder to recognize than a normal face.

Hair, eyes, mouth, face outline have been determined to be more important than nose for perceiving and remembering faces. It has also been found that the upper part of the face is more useful than the lower part of the face for recognition. Also, aesthetic attributes (e.g. beauty, attractiveness, pleasantness, etc.) play an important role in face recognition; the more attractive the faces are easily remembered.

For humans, photographic negatives of faces are difficult to recognize. But, there is not much study on why it is difficult to recognize negative images of human faces. Also, a study on the direction of illumination [7] showed the importance of top lighting; it is easier for humans to recognize faces illuminated from top to bottom than the faces illuminated from bottom to top.

According to the neurophysicists, the analysis of facial expressions is done in parallel to face recognition in human face recogniton system. Some prosopagnosic patients, who have difficulties in identifying familiar faces, seem to recognize facial expressions due to emotions. Patients who suffer from organic brain syndrome do poorly at expression analysis but perform face recognition quite well.

2.2. Machine Recognition of Faces

Although studies on human face recognition were expected to be a reference on machine recognition of faces, research on machine recognition of faces has developed independent of studies on human face recognition. During 1970's, typical pattern classification techniques, which use measurements between features in faces or face profiles, were used [4]. During the 1980's, work on face recognition remained nearly

stable. Since the early 1990's, research interest on machine recognition of faces has grown tremendously. The reasons may be;

- An increase in emphasis on civilian/commercial research projects,
- The studies on neural network classifiers with emphasis on real-time computation and adaptation,
- The availibility of real time hardware,
- The growing need for surveillence applications.

The basic question relevant for face classification is that; what form the structural code (for encoding the face) should take to achieve face recognition. Two major approaches are used for machine identification of human faces; geometrical local feature based methods, and holistic template matching based systems. Also, combinations of these two methods, namely hybrid methods, are used. The first approach, the geometrical local feature based one, extracts and measures discrete local features (such as eye, nose, mouth, hair, etc.) for retrieving and identifying faces. Then, standard statistical pattern recognition techniques and/or neural network approaches are employed for matching faces using these measurements [8]. One of the well known geometrical-local feature based methods is the Elastic Bunch Graph Matching (EBGM) technique. The other approach, the holistic one, conceptually related to template matching, attempts to identify faces using global representations [2]. Holistic methods approach the face image as a whole and try to extract features from the whole face region. In this approach, as in the previous approach, the pattern classifiers are applied to classify the image after extracting the features. One of the methods to extract features in a holistic system is applying statistical methods such as Principal

Component Analysis (PCA) to the whole image. PCA can also be applied to a face image locally; in that case the approach is not holistic.

Whichever method is used, the most important problem in face recognition is the curse of dimensionality problem. Appropriate methods should be applied to reduce the dimension of the studied space. Working on higher dimension causes overfitting where the system starts to memorize. Also, computational complexity would be an important problem when working on large databases.

In the following sections, the main studies shall be summarized. The recognition techniques are grouped as statistical and neural based approaches.

2.2.1. Statistical Approaches

Statistical methods include template matching based systems where the training and test images are matched by measuring the correlation between them. Moreover, statistical methods include the projection based methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), etc. In fact, projection based systems came out due to the shortcomings of the straightforward template matching based approaches; that is, trying to carry out the required classification task in a space of extremely high dimensionality.

- *Template Matching:* Brunelli and Poggio [9] suggest that the optimal strategy for face recognition is holistic and corresponds to template matching. In their study, they compared a geometric feature based technique with a template matching based system. In the simplest form of template matching, the image (as 2-D intensity values) is

compared with a single template representing the whole face using a distance metric.

Although recognition by matching raw images has been successful under limited circumstances, it suffers from the usual shortcomings of straightforward correlation-based approaches, such as sensitivity to face orientation, size, variable lighting conditions, and noise. The reason for this vulnerability of direct matching methods lies in their attempt to carry out the required classification in a space of extremely high dimensionality. In order to overcome the curse of dimensionality, the connectionist equivalent of data compression methods is employed first. However, it has been successfully argued that the resulting feature dimensions do not necessarily retain the structure needed for classification, and that more general and powerful methods for feature extraction such as projection based systems are required. The basic idea behind projection based systems is to construct low dimensional projections of a high dimensional point cloud, by maximizing an objective function such as the deviation from normality.

2.2.1.1. Face Detection and Recognition by PCA

The Eigenface Method of Turk and Pentland [10] is one of the main methods applied in the literature which is based on the Karhunen-Loeve expansion. Their study is motivated by the earlier work of Sirowich and Kirby [11], [12]. It is based on the application of Principal Component Analysis to the human faces. It treats the face images as 2-D data, and classifies the face images by projecting them to the eigenface space which is composed of eigenvectors obtained by the variance of the face images. Eigenface recognition derives its name from the German prefix *eigen*, meaning own or individual. The Eigenface method of facial

recognition is considered the first working facial recognition technology [13].

When the method was first proposed by Turk and Pentland [10], they worked on the image as a whole. Also, they used Nearest Mean classifier two classify the face images. By using the observation that the projection of a face image and non-face image are quite different, a method of detecting the face in an image is obtained. They applied the method on a database of 2500 face images of 16 subjects, digitized at all combinations of 3 head orientations, 3 head sizes and 3 lighting conditions. They conducted several experiments to test the robustness of their approach to illumination changes, variations in size, head orientation, and the differences between training and test conditions. They reported that the system was fairly robust to illumination changes, but degrades quickly as the scale changes [10]. This can be explained by the correlation between images obtained under different illumination conditions; the correlation between face images at different scales is rather low. The eigenface approach works well as long as the test image is similar to the training images used for obtaining the eigenfaces.

Later, derivations of the original PCA approach are proposed for different applications.

- *PCA and Image Compression:* In their study Moghaddam and Pentland [14] used the Eigenface Method for image coding of human faces for potential applications such as video telephony, database image compression and face recognition.

- *Face Detection and Recognition Using PCA:* Lee et al. [15] proposed a method using PCA which detects the head of an individual in a complex background and then recognize the person by comparing the characteristics of the face to those of known individuals.

- *PCA Performance on Large Databases:* Lee et al. [16] proposed a method for generalizing the representational capacity of available face database.

- *PCA & Video:* In a study by Crowley and Schwerdt [17], PCA is used for coding and compression for video streams of talking heads. They suggest that a typical video sequence of a talking head can often be coded in less than 16 dimensions.

- *Bayesian PCA:* Another method, which is also studied throughout this thesis, is the Bayesian PCA method suggested by Moghaddam et al. [24], [25], [26], [27], [28], [29], [30], [31]. By this system, the Eigenface Method based on simple subspace-restricted norms is extended to use a probabilistic measure of similarity. Also, another difference from the standard Eigenface approach is that this method uses the image differences in the training and test stages. The difference of each image belonging to the same individual with each other is fed into the system as intrapersonal difference, and the difference of one image with an image from different class is fed into the system as extrapersonal difference. Finally, when a test image comes, it is subtracted from each image in the database and each difference is fed into the system. For the biggest similarity (i.e. smallest difference) with one of the training images, the test image is decided to be in that class. The mathematical theory is mainly studied in [32]

Also, in [33] Moghaddam introduced his study on several techniques; Principal Component Analysis (PCA), Independent Component Analysis (ICA), and nonlinear Kernel PCA (KPCA). He examined and tested these systems using the FERET database. He argued that the experimental results demonstrate the simplicity, computational

economy and performance superiority of the Bayesian PCA method over other methods.

Finally, Liu and Wechsler [34], [35] worked on Bayesian approach to face recognition.

- *PCA and Gabor Filters:* Chung et al. [36] suggested the use of PCA and Gabor Filters together. Their method consists of two parts: In the first part, Gabor Filters are used to extract facial features from the original image on predefined fiducial points. In the second part, PCA is used to classify the facial features optimally. They suggest the use of combining these two methods in order to overcome the shortcomings of PCA. They argue that, when raw images are used as a matrix of PCA, the eigenspace cannot reflect the correlation of facial feature well, as original face images have deformation due to in-plane, in-depth rotation and illumination and contrast variation. Also they argue that, they have overcome these problems using Gabor Filters in extracting facial features.

2.2.1.2. Face Recognition by LDA

Etemad and Chellappa [19] proposed a method on appliance of Linear/Fisher Discriminant Analysis for the face recognition process. LDA is carried out via scatter matrix analysis. The aim is to find the optimal projection which maximizes between class scatter of the face data and minimizes within class scatter of the face data. As in the case of PCA, where the eigenfaces are calculated by the eigenvalue analysis, the projections of LDA are calculated by the generalized eigenvalue equation.

- *Subspace LDA:* An alternative method which combines PCA and LDA is studied [20], [21], [22], [23]. This method consists of two

steps; the face image is projected into the eigenface space which is constructed by PCA, and then the eigenface space projected vectors are projected into the LDA classification space to construct a linear classifier. In this method, the choice of the number of eigenfaces used for the first step is critical; the choice enables the system to generate class-separable features via LDA from the eigenface space representation. The generalization/overfitting problem can be solved in this manner. In these studies, a weighted distance metric guided by the LDA eigenvalues was also employed to improve the performance of the subspace LDA method.

2.2.1.3. Transformation Based Systems

- *DCT:* Podilchuk and Zang [37] proposed a method which finds the feature vectors using Discrete Cosine Transform (DCT). Their system tries to detect the critical areas of the face. The system is based on matching the image to a map of invariant facial attributes associated with specific areas of the face. They claim that this technique is quite robust, since it relies on global operations over a whole region of the face. A codebook of feature vectors or codewords is determined for each person from the training set. They examine recognition performance based on feature selection, number of features or codebook size, and feature dimensionality. For feature selection, they tried several block-based transformations and the K-means clustering algorithm [44] to generate the codewords for each codebook. They argue that the block-based DCT coefficients produce good low-dimensional feature vectors with high recognition performance. This brings the possibility of performing face recognition directly on a DCT-based compressed bitstream without having to decode the image.

- *DCT & HMMs:* Eickeler et al. [38] suggested a system based on Pseudo 2-D Hidden Markow Madels (HMMs) and coefficient of the 2-D

DCT as the features. A major advantage of their approach is that the system works directly on JPEG-compressed face images, i.e. it uses the DCT-coefficients provided by the JPEG standard. Thus, it does not need any further decompressing of the image. Also Nefian and Hayes [39] studied on first using DCT as the feature vectors and using HMMs.

- *Fourier Transform (FT):* Spies and Ricketts [41] describe a face recognition system based on an analysis of faces via their Fourier spectra. Recognition is achieved by finding the closest match between feature vectors containing the Fourier coefficients at selected frequencies. This technique is based on the Fourier spectra of facial images, thus it relies on a global transformation, i.e. every pixel in the image contributes to each value of its spectrum. The Fourier spectrum is a plot of the energy against spatial frequencies, where spatial frequencies relate to the spatial relations of intensities in the image. In the case of face recognition, this translates to distances between areas of particular brightness, such as the overall size of the head, or the distance of the eyes. Higher frequencies describe finer details and they claim that these are less useful for identification of a person. They also suggest that, humans can recognize a face from a brief look without focusing on small details. They perform the recognition of faces by finding the Euclidian distance between a newly presented face and all the training faces. The distances are calculated between feature vectors with entries that are the Fourier Transform values at specially chosen frequencies. They argue that, as few as 27 frequencies yield good results (98 %). Moreover, this small feature vector combined with the efficient Fast Fourier Transform (FFT) makes this system extremely fast.

2.2.1.4. Face Recognition by SVM

Phillips [40] applied SVM to face recognition. Face recognition is a K-class problem, where K is the number of known individuals; and SVM is a binary classification method. By reformulating the face recognition problem and reinterpreting the output of the SVM classifier, they developed a SVM-based face recognition algorithm. They formulated the face recognition problem in difference space, which models dissimilarities between two facial images. In difference space, they formulated the face recognition as a two class problem. The classes are; dissimilarities between faces of the same person and dissimilarities between faces of different people. By modifying the interpretation of the decision surface generated by SVM, they generated a similarity metric between faces, learned from examples of differences between faces [40].

2.2.1.5. Feature-based Approaches

Bobis et al. [42] studied on a feature based face recognition system. They suggested that a face can be recognized by extracting the relative position and other parameters of distinctive features such as eyes, mouth, nose and chin. The system described the overall geometrical configuration of face features by a vector of numerical data representing position and size of main facial features. First, they extracted eyes coordinates. The interocular distance and eyes position is used to determine size and position of the areas of search for face features. In these areas binary thresholding is performed, system modifies threshold automatically to detect features. In order to find their coordinates, discontinuities are searched for in the binary image. They claim that, their experimental results showed that their method is robust, valid for numerous kind of facial image in real scene, works in real time with low hardware requirements and the whole process is conducted automatically.

- *Feature Based PCA:* Cagnoni and Poggi [18] suggested a feature-based approach instead of a holistic approach to face recognition. They applied the eigenface method to sub-images (eye, nose, and mouth). They also applied a rotation correction to the faces in order to obtain better results.

- *A feature based approach using PCA vs. ICA:* Guan and Szu [43] compared the performance of PCA and ICA on face images. They argue that, ICA encodes face images with statistically independent variables, which are not necessarily associated with the orthogonal axes, while PCA is always associated with orthogonal eigenvectors. While PCA seeks directions in feature space that best represent the data in a sum-squared error sense, ICA seeks directions that are most independent from each other [44]. They also argue that, both these pixel-based algorithms have the major drawback that they weight the whole face equally and therefore lack the local geometry information. Hence, Guan and Szu suggest approaching the face recognition problem with ICA or PCA applied on local features.

- *A feature based approach using PCA, Bayesian Classifier, and HMMs:* Martinez [45] proposed a different approach based on identifying frontal faces. Their approach divides a face image into n different regions, analyzes each region with PCA, and then uses a Bayesian approach to find the best possible global match between a probe and database image. The relationship between the n parts is modeled by using Hidden Markov Models (HMMs).

2.2.2. Neural Network Approaches

Neural Network approaches have been used in face recognition generally in a geometrical local feature based manner, but there are also some methods where neural networks are applied holistically.

- *Feature based Backpropagation NN:* Temdee et al. [46] presented a frontal view face recognition method by using fractal codes which are determined by a fractal encoding method from the edge pattern of the face region (covering eyebrows, eyes and nose). In their recognition system, the obtained fractal codes are fed as inputs to a Backpropagation Neural Network for identifying an individual. They tested their system performance on the ORL face database. They report their performance as 85 % correct recognition rate in the ORL face database.

- *Dynamic Link Architectures (DLA):* Lades et al. [47] presented an object recognition system based on Dynamic Link Architectures, which is an extension of the Artificial Neural Networks. The DLA uses correlations in the fine-scale cellular signals to group neurons dynamically into higher order entities. These entities can be used to code high-level objects, such as a 2-D face image. The face images are represented by sparse graphs, whose vertices are labeled by a multiresolution description in terms of local power spectrum, and whose edges are labeled by geometrical distance vectors. Face recognition can be formulated as elastic graph matching, which is performed in this study by stochastic optimization of a matching cost function.

- *Elastic Bunch Graph Matching (EBGM):* Wiskott et al. [48] presented a geometrical local feature based system for face recognition from single images out of a large database containing one image per person, which is known as Elastic Bunch Graph Matching (EBGM). In

this system, faces are represented by labeled graphs, based on a Gabor Wavelet Transform (GWT). Image graphs of new faces are extracted by an Elastic Graph Matching process and can be compared by a simple similarity function. In this system, phase information is used for accurate node positioning and object-adapted graphs are used to handle large rotations in depth. The image graph extraction is based on the bunch graph, which is constructed from a small set of sample image graphs. In contrast to many neural-network systems, no extensive training for new faces or new object classes is required. Only a small number of typical examples have to be inspected to build up a bunch graph, and individuals can then be recognized after storing a single image.

The system inhibits most of the variance caused by position, size, expression and pose changes by extracting concise face descriptors in the form of image graphs. In these image graphs, some predetermined points on the face (eyes, nose, mouth, etc.) are described by sets of wavelet components (jets). The image graph extraction is based on the bunch graph, which is constructed from a small set of image graphs.

2.2.3. Hybrid Approaches

There are some other approaches which use both statistical pattern recognition techniques and Neural Network systems.

- *PCA and RBF:* The method by Er et al. [49] suggests the use of Radial Basis Function (RBF) Neural Networks on the data extracted by discriminant eigenfeatures. They used a hybrid learning algorithm to decrease the dimension of the search space in the gradient method, which is crucial on optimization of high dimension problem. First, they tried to extract the face features by both the PCA and LDA methods. Next, they presented a hybrid learning algorithm to train the RBF Neural Networks,

so the dimension of the search space is significantly decreased in the gradient method.

Thomaz et al. [50] also studied on combining PCA and RBF neural network. Their system is a face recognition system consisting of a PCA stage which inputs the projections of a face image over the principal components into a RBF network acting as a classifier. Their main concern is to analyze how different network designs perform in a PCA+RBF face recognition system. They used a forward selection algorithm, and a Gaussian mixture model. According to the results of their experiments, the Gaussian mixture model optimization achieves the best performance even using less neurons than the forward selection algorithm. Their results also show that the Gaussian mixture model design is less sensitive to the choice of the training set.

2.2.4. Other Approaches

2.2.4.1. Range Data

One of the different methods used in face recognition task is using the range images. In this method data is obtained by scanning the individual with a laser scanner system. This system also has the depth information so the system processes 3-dimensonal data to classify face images [4].

2.2.4.2. Infrared Scanning

Another method used for face recognition is scanning the face image by an infrared light source.

Yoshitomi et al. [51] used thermal sensors to detect temperature distribution of a face. In this method, the front-view face in input image is normalized in terms of location and size, followed by measuring the

temperature distribution, the locally averaged temperature and the shape factors of face. The measured temperature distribution and the locally averaged temperature are separately used as input data to feed a Neural Network, while the values of shape factors are used for supervised classification. By integrating information from the Neural Network and supervised classification, the face is identified. The disadvantage of visible ray image analysis that the accuracy of face identification is strongly influenced by lighting condition including variation of shadow, reflection and darkness is overcome by this method which uses infrared rays.

2.2.4.3. Profile Images

Liposcak and Loncaric [52] worked on profile images instead of frontal images. Their method is based on the representation of the original and morphological derived profile shapes. Their aim is to use the profile outline that bounds the face and the hair. They take a grey-level profile image, threshold it to produce a binary image, representing the face region. They normalize the area and orientation of this shape using dilation and erosion. Then, they simulate hair growth and haircut and produce two new profile silhouettes. From these three profile shapes they obtain the feature vectors. After normalizing the vector components, they use the Euclidean distance measure for measuring the similarity of the feature vectors derived from different profiles.

CHAPTER 3

PATTERN CLASSIFICATION

In this chapter, the idea behind the algorithms applied throughout the thesis study will be summarized. The pattern classification basics, the discipline behind the statistical pattern classifiers; Bayesian Classifier, Nearest Mean Classifier, and the main projection techniques; Principal Component Analysis and Linear Discriminant Analysis shall be explained.

3.1. Pattern Classification Basics

Pattern recognition is the science that concerns the description or classification of measurements. A typical pattern recognition system takes the input data, extracts the features, trains the classifier and evaluates the test pattern. The features are the representatives of the input data in the classification system used and they are obtained by directly using the input data or by applying different kinds of dimension reduction techniques on the input data. An ideal feature extractor would yield a representation that makes the job of the classifier trivial.

There are three main pattern recognition approaches [53];

- Statistical Pattern Recognition; which assumes the underlying model as a set of probabilities, but ignores the structure,

- Syntactic Pattern Recognition; which concentrates on the interrelations between primitives that build the whole pattern (which are not easy to find),
- Neural Pattern Recognition; which imitates human neural system.

Throughout this thesis study, statistical pattern classification is used. Hence, it will be the main concern to discuss in this chapter.

Statistical pattern recognition is an approach of the pattern classification problem, which concentrates on developing decision or classification strategies that form classifiers [53]. In statistical pattern recognition, each pattern is represented in terms of d features or measurements and is can be viewed as a point in a d-dimensional space. The objective of a statistical pattern recognition system is to choose features that allow the patterns belonging to different classes to occupy different regions in a d-dimensional feature space. The representation space or feature space is assumed to be effective, if the patterns from different classes are well separated and if the feature space greatly represents the general properties of the input data. Decision boundaries separate patterns belonging to different classes and they are determined by the specified probability distributions of the patterns belonging to each class.

While a heavily complex system may allow perfect classification of the training samples, it may not perform well on new patterns. This situation is known as overfitting (or memorization) [44]. One of the most important areas of research in statistical pattern classification is determining how to adjust the complexity of the model.

For the case where the training samples used to design a classifier are labeled by their category membership, the procedure is accepted as *supervised*. The opposite case is called *unsupervised* system where the training data is unlabeled.

3.2. Classifiers

3.2.1. Bayesian Classifier

Bayesian classifier is one of the most widely applied statistical approaches in the standard pattern classification. It assumes that the classification problem is posed in probabilistic terms, and that all the relevant probability values are known.

The decision process in statistical pattern recognition can be explained as follows: Given an input pattern, represented as a vector of d feature values $x=(x_1, x_2, \ldots, x_d)$, assign it to one of c categories $w_1, w_2, \ldots, w_c$ by taking one of the a possible actions $\alpha_1, \alpha_2, \ldots, \alpha_a$. Assuming that each feature has a probability density (represented as $p(.)$) or mass function (represented as $P(.)$) conditioned on the pattern class, a pattern vector x belonging to class w_i can be viewed as an observation drawn randomly from the class-conditional probability function, $p(x|w_i)$ [54].

A loss function $\lambda(\alpha_i|w_j)$ is simply defined as the loss for deciding the class w_i where the true class is class w_j [44], that is;

Loss Function
$$\lambda(\alpha_i|w_j)=\begin{cases}0, & i=j\\ 1, & i\neq j\end{cases} \tag{3.1}$$

The Bayes decision rule for minimizing the risk, which is the expected value of the loss function, can be stated as follows [44]: Assign input pattern x to class w_i where the conditional risk $R(\alpha_i|x)$ is minimum. In mathematical terms;

Conditional Risk $$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j) P(w_j|x) \quad (3.2)$$

should be minimized for any i = 1, 2, …, c to classify the input pattern to the i^{th} class.

Under these assumptions, the Bayes classifier assigns the input pattern x to class w_i for which the posterior probability $P(w_i|x)$ is the maximum; that is,

Bayes Classifier Decide on w_i if $P(w_i|x) > P(w_j|x)$ for $i \neq j$ (3.3)

The idea underlying Bayes classifier is very simple. In order to minimize the overall risk, the action that minimizes the conditional risk $R(\alpha|x)$ should be taken. In particular, in order to minimize the probability of error in a classification problem, the maximum posterior probability $P(w_i|x)$ should be chosen. Bayes formula is a way to calculate such probabilities from the prior probabilities $P(w_j)$ and the class conditional densities $p(x|w_j)$[44].

Bayes Formula $$P(w_j|x) = \frac{p(x|w_j)P(w_j)}{p(x)} = \frac{p(x|w_j)P(w_j)}{\sum_{j=1}^{c} p(x|w_j)P(w_j)} \quad (3.4)$$

For most of the pattern classification applications, the main problem in applying the Bayes classifier is that the class-conditional

densities $p(x|w_j)$ are not known. However, in some cases, the form of these densities can be known without exact knowledge on parameter values. In a classic case, the densities are known to be (or assumed to be) multivariate normal, but the values of the mean vector and the covariance matrix are not known [44]. In this case, the multivariate normal density in *d* dimensions can be written as [34];

$$\textit{Multivariate Normal Density} \quad p(x|w_j) = \frac{e^{\left(-\frac{1}{2}(x-\mu_j)^T \Sigma^{-1}(x-\mu_j)\right)}}{(2\pi)^{d/2}|\Sigma|^{1/2}} \qquad (3.5)$$

where Σ is the covariance matrix of size (*d* x *d*).

3.2.2. Nearest Mean Classifier

Nearest Mean Classifier is an analogous approach to the Nearest Neighbor Rule (NN-Rule). In the NN-Rule, after the classification system is trained by the samples, a test data is fed into the system and it is classified in the class of the nearest training sample in the data space with respect to Euclidean Distance. In the Nearest Mean Classifier, again the Euclidean distance from each class mean (in this case) is computed for the decision of the class of the test data [54].

In mathematical terms, the Euclidean distance between the test sample *x*, and each face class mean μ_i is;

$$\textit{Euclidean Distance} \qquad d_i = \|x - \mu_i\| = \sqrt{\sum_{k=1}^{d} (x_k - \mu_{ik})^2} \qquad (3.6)$$

where *x* is an *d* dimensional input data.

After computing the distance to each class mean, the test data is classified into the class with minimum Euclidean Distance.

3.3. Feature Dimension Reduction

Feature selection in pattern recognition involves the derivation of certain features from the input data in order to reduce the amount of data used for classification and provide discrimination power. Due to the measurement cost and classification accuracy, the number of features should be kept as small as possible. A small and functional feature set makes the system work faster and use less memory. On the other hand, using a wide feature set, may cause "curse of dimensionality" which is the need for exponentially growing number of samples [44].

Feature extraction methods try to reduce the feature dimensions used in the classification step. There are especially two methods used in pattern recognition to reduce the feature dimensions; Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [44].

The relative places of the data between each other is never changed according to the utilized feature dimension reduction technique, PCA or LDA. Only the axes are changed to handle the data from a "better" point of view. Better point of view is simply generalization for PCA and discrimination for LDA.

3.3.1. Principal Component Analysis (PCA)

The advantage of PCA comes from its generalization ability. It reduces the feature space dimension by considering the variance of the input data. The method determines which projections are preferable for representing the structure of the input data. Those projections are selected in such a way that the maximum amount of information (i.e. maximum

variance) is obtained in the smallest number of dimensions of feature space.

In order to obtain the best variance in the data, the data is projected to a subspace (of the image space) which is built by the eigenvectors from the data. In that sense, the eigenvalue corresponding to an eigenvector represents the amount of variance that eigenvector handles. The mathematical formulation of PCA is discussed in Chapter 4.1.

3.3.2. Linear Discriminant Analysis (LDA)

While PCA tries to generalize the input data to extract the features, LDA tries to discriminate the input data by dimension reduction.

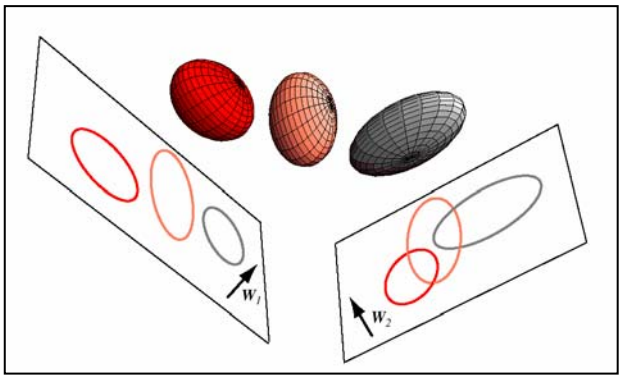


Figure 3.1: Three 3-dimensional distributions are projected onto 2-dimensional subspaces, described by normal vectors $\mathbf{W_1}$ and $\mathbf{W_2}$. As LDA tries to find the greatest seperation among the classes, $\mathbf{W_1}$ is optimal for LDA [44].

LDA searches for the best projection to project the input data, on a lower dimensional space, in which the patterns are discriminated as much as possible. For this purpose, LDA tries to maximize the scatter [44] between different classes and minimize the scatter between the input datas in the same class. In Figure 3.1, the selection criteria for LDA can be observed. LDA uses generalized eigenvalue equation [44] to obtain this discrimination. The mathematical aspects of the LDA can be found in Chapter 4.2.

CHAPTER 4

FACE RECOGNITION VIA DIMENSION REDUCTION

In this chapter, the theoretical background of the implemented methods will be discussed. These are namely Eigenface Method (PCA) of Turk and Pentland [10], pure LDA [44], Subspace LDA (PCA + LDA) of Zhao et al. [21], [22], [23], and Bayesian PCA (PCA + Bayes Classifier) of Moghaddam et al. [24], [25], [26], [27], [28], [29], [30], [31].

In Subspace LDA, PCA is used for the purpose of dimension reduction by generalizing the data and LDA is used for classification due to its discrimination power. Also in Bayesian PCA, PCA is used for the purpose of dimension reduction by generalizing the data and Bayes Classifier is used for its good data modeling power.

After this brief introduction, in the next sections Eigenface, pure LDA, Subspace LDA and Bayesian PCA methods shall be explained in detail.

4.1. Eigenface Method

An image space can be thought of as a space having dimensions equal to the number of pixels making up the image and having values in

the range of the pixels values. Thus, for example for a grey scale image of size (N_x x N_y), the dimension of the image space is P, P being N_x times N_y. For the case of gray scale images, in each dimension the image could have a value in between 0 and 255.

An image can be thought as a point in the image space by converting the image to a long vector by concatenating each column of the image one after the other.

When all the face images are converted into vectors, they will group at a certain location in the image space as they have similar structure, having eye, nose and mouth in common and their relative position correlated. This correlation is the main point to start the eigenface analysis.

The Eigenface method tries to find a lower dimensional space for the representation of the face images by eliminating the variance due to non-face images; that is, it tries to focus on the variation just coming out of the variation between the face images.

Eigenface method is the implementation of Principal Component Analysis (PCA) over images. In this method, the features of the studied images are obtained by looking for the maximum deviation of each image from the mean image. This variance is obtained by getting the eigenvectors of the covariance matrix of all the images.

The eigenface space is obtained by applying the eigenface method to the training images. Later, the training images are projected into the eigenface space. Next, the test image is projected into this new space and the distance of the projected test image to the training images is used to classify the test image. In the standard eigenface procedure suggested by

Turk and Pentland [10], Nearest Mean Classifier is used for the classification of test images.

In mathematical terms;

Image $\quad I : \quad (N_x \text{ x } N_y) \text{ pixels} \quad (4.1)$

The image matrix I of size (N_x x N_y) pixels is converted to the image vector Γ of size (P x 1) where P = (N_x x N_y); that is the image matrix is reconstructed by adding each column one after the other.

Training Set $\quad \Gamma = [\Gamma_1 \quad \Gamma_2 \quad ... \quad \Gamma_{Mt}] \quad (4.2)$

is the training set of image vectors and its size is (P x M_t) where M_t is the number of the training images.

Mean Face $\quad \Psi = \frac{1}{M_t} \sum_{i=1}^{M_t} \Gamma_i \quad (4.3)$

is the arithmetic average of the training image vectors at each pixel point and its size is (P x 1).

Mean subtracted image $\quad \Phi = \Gamma - \Psi \quad (4.4)$

is the difference of the training image from the mean image (size P x 1).

Difference Matrix $\quad A = [\Phi_1 \quad \Phi_2 \quad ... \quad \Phi_{Mt}] \quad (4.5)$

is the matrix of all the mean subtracted training image vectors and its size is (P x M_t).

Covariance Matrix $\quad X = A \cdot A^T = \frac{1}{M_t} \sum_{i=1}^{M_t} \Phi_i \Phi_i^T \quad (4.6)$

is the covariance matrix of the training image vectors of size (P x P).

An important property of the Eigenface method is obtaining the eigenvectors of the covariance matrix. For a face image of size (N_x x N_y)

pixels, the covariance matrix is of size (P x P), P being (N_x x N_y). This covariance matrix is very hard to work with due to its huge dimension causing computational complexity. On the other hand, Eigenface method calculates the eigenvectors of the (M_t x M_t) matrix, M_t being the number of face images, and obtains (P x P) matrix using the eigenvectors of the (M_t x M_t) matrix.

Initially, a matrix Y is defined as,

$$\mathrm{Y} = \mathrm{A}^T \cdot \mathrm{A} = \frac{1}{M_t} \sum_{i=1}^{M_t} \Gamma_i^T \Gamma_i \tag{4.7}$$

which is of size (M_t x M_t).

Then, the eigenvectors v_i and the eigenvalues μ_i of Y are obtained,

$$\mathrm{Y} \cdot v_i = \mu_i \cdot v_i \tag{4.8}$$

The value of Y is put in this equation,

$$\mathrm{A}^T \cdot \mathrm{A} \cdot v_i = \mu_i \cdot v_i \tag{4.9}$$

Both sides are left multiplied by A,

$$\mathrm{A} \cdot \mathrm{A}^T \cdot \mathrm{A} \cdot v_i = \mathrm{A} \cdot \mu_i \cdot v_i \tag{4.10}$$

The necessary matrix arrangements are made,

$$\mathrm{A} \cdot \mathrm{A}^T \cdot \mathrm{A} \cdot v_i = \mu_i \cdot \mathrm{A} \cdot v_i \tag{4.11}$$

(as μ_i is a scalar, this arrangement can be done)

$$\mathrm{X} \cdot \mathrm{A} \cdot v_i = \mu_i \cdot \mathrm{A} \cdot v_i \tag{4.12}$$

Now group $\mathrm{A} \cdot v_i$ and call a variable $\upsilon_i = \mathrm{A} \cdot v_i$. It is easy to see that

$$\upsilon_i = \mathrm{A} \cdot v_i \tag{4.13}$$

is one of the eigenvectors of $X = A \cdot A^T$ and its size is (P x 1).

Thus, it is possible to obtain the eigenvectors of X by using the eigenvectors of Y. A matrix of size (M_t x M_t) is utilized instead of a matrix of size (P x P) (i.e. [{N_x x N_y} x {N_x x N_y}]). This formulation brings substantial computational efficiency. In Figure 4.1, some example images and mean image of the images from the ORL database (shall be detailed in Chapter 5.1.1) are given. In Figure 4.2, some characteristic eigenfaces obtained from this database can be seen. The eigenfaces are in fact (P x 1) vectors for the computations; in order to see what they look like, they are rearranged as (N_x x N_y) matrices.

(a)

(b)

Figure 4.1: (a) Example images from the ORL database, **(b)** Mean face obtained from the ORL database

Instead of using M_t of the eigenfaces, $M' \leq M_t$ of the eigenfaces can be used for the eigenface projection. This is achieved to eliminate some of the eigenvectors with small eigenvalues, which contribute less variance in the data. In Figure 4.3, the cumulative sum of the eigenvalues and in Figure 4.4, a typical eigenvalue spectrum can be observed.

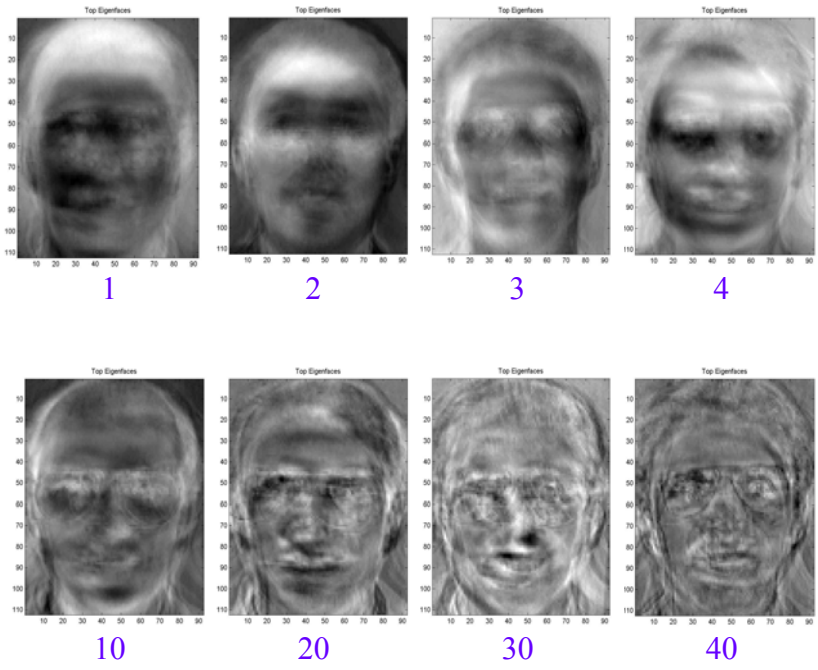


Figure 4.2: Some examples of the eigenfaces, sorted with respect to decreasing eigenvalues.

From the Figures 4.3 and 4.4, it can be easily observed that most of the generalization power is contained in the first few eigenvectors. For example, 40 % of the total eigenvectors have 85 – 90 % of the total generalization power. Thus, using 40 % of the total number of eigenvectors may end up with reasonable classification results.

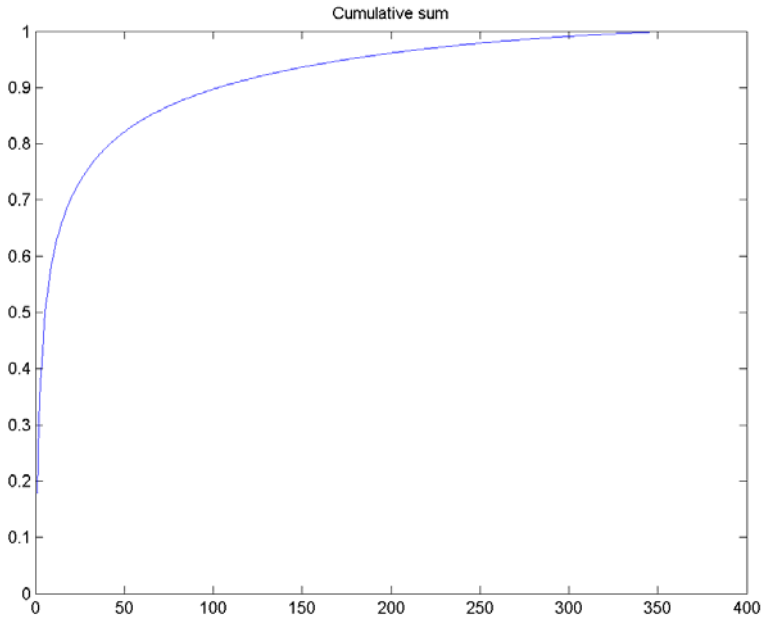


Figure 4.3: The cumulative sum curve for the eigenvalues

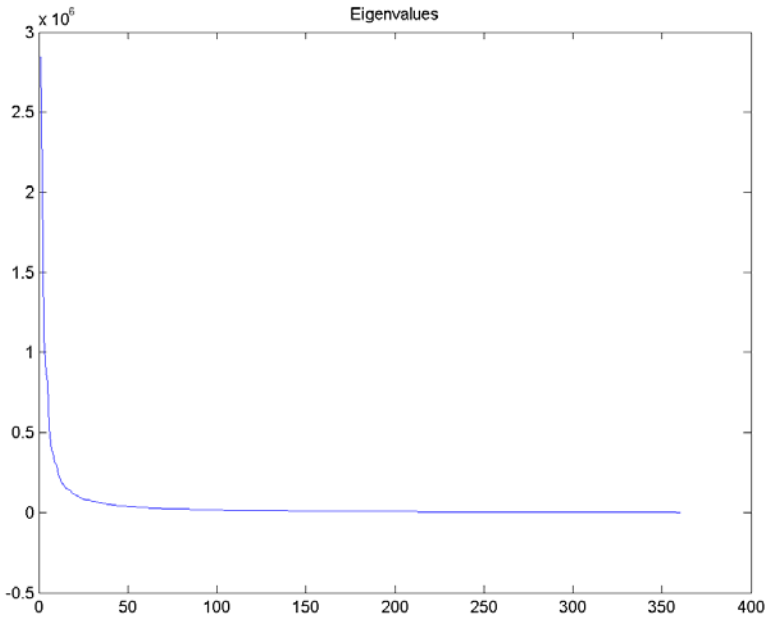


Figure 4.4: A typical eigenvalue spectrum

Eigenvectors can be considered as the vectors pointing in the direction of the maximum variance and the value of the variance the eigenvector represents is directly proportional to the value of the eigenvalue (i.e. the larger the eigenvalue indicates the larger variance the eigenvector represents). Hence, the eigenvectors are sorted with respect to their corresponding eigenvalues. The eigenvector having the largest eigenvalue is marked as the first eigenvector, and so on. In this manner, the most generalizing eigenvector comes first in the eigenvector matrix.

In the next step, the training images are projected into the eigenface space and thus the weight of each eigenvector to represent the image in the eigenface space is calculated. This weight is simply the dot product of each image with each of the eigenvectors.

Projection $$\omega_k = \upsilon_k^T \cdot \Phi = \upsilon_k^T \cdot (\Gamma - \Psi) \tag{4.14}$$

is the projection of a training image on each of the eigenvectors where k = 1, 2, …, M′

Weight Matrix $$\Omega = \left[\omega_1 \quad \omega_2 \quad \ldots \quad \omega_{M'}\right]^T \tag{4.15}$$

is the representation of the training image in the eigenface space and its size is (M′ x 1).

At this point, the images are just composed of weights in the eigenface space, simply like they have pixel values in the image space. The important aspect of the eigenface transform lies in this property. Each image is represented by an image of size (N_x x N_y) in the image space, whereas the same image is represented by a vector of size (M′ x 1) in the eigenface space. Moreover, having the dimension structure related to the variance of the data in hand makes the eigenface representation a generalized representation of the data. This makes the algorithm a

solution to the "curse of dimensionality" problem [44] seen in the standard pattern recognition task.

When a new test image is to be classified, it is also mean subtracted and projected onto the eigenface space and Nearest Mean algorithm [10] is used for the classification of the test image vector in the standard eigenface method; that is, the test image is assumed to belong to the nearest class by calculating the Euclidean distance of the test image vector to the mean of each class of the training image vectors.

Test image vector Γ_T (4.16)

is the test image vector of size (P x 1).

Mean subtracted image $\Phi_T = \Gamma_T - \Psi$ (4.17)

is the difference of the test image from the mean image (size P x 1).

Projection $\omega_k = \upsilon_k^T \cdot \Phi_T = \upsilon_k^T \cdot (\Gamma_T - \Psi)$ (4.18)

is the projection of a training image on each of the eigenvectors where k = 1, 2, …, M′

Weight Matrix $\Omega_T = \left[\omega_1 \quad \omega_2 \quad \ldots \quad \omega_{M'}\right]^T$ (4.19)

is the representation of the test image in the eigenface space and its size is (M′ x 1).

The training and test image vectors can be reconstructed by a back transformation from the eigenface space to the image vector space.

Reconstructed image vector $\Gamma_f = \upsilon \cdot \Omega + \Psi = \Phi_f + \Psi$ (4.20)

If possible, it is better to work on a database of more than one image per individual in order to increase the robustness to minor changes in expression, illumination and slight variations of view angles. A class

of images for an individual can be formed and this class can be considered as the representative image vector of that class. For an individual having q_i images in the database, the average of the projections of each class is the mean of all the projected image vectors in that class. In mathematical terms;

$$\textit{Average Class Projection} \qquad \Omega_\Psi = \frac{1}{q_i}\sum_{i=1}^{q_i}\Omega_i \tag{4.21}$$

This average class projection can be used as one of the vectors (representing an image class instead of an image vector) to compare with the test image vector.

A similarity measure is defined as the distance between the test image vector and i-th face class;

$$\textit{Similarity} \qquad \delta_i = \left\|\Omega_T - \Omega_{\Psi i}\right\| = \sqrt{\sum_{k=1}^{Mt}\left(\Omega_{Tk} - \Omega_{\Psi ik}\right)^2} \tag{4.22}$$

is the Euclidean distance (L2 norm) between projections.

A distance threshold may be defined for the maximum allowable distance from any face class, which is half of the distance between the two most distant classes;

$$\textit{Distance Threshold} \qquad \Theta = \frac{1}{2}\max\left(\left\|\Omega_{\Psi i} - \Omega_{\Psi j}\right\|\right) \tag{4.23}$$

Classification procedure of the Eigenface method ensures that face image vectors should fall close to their reconstructions, whereas non-face image vectors should fall far away. Hence, a distance measure is defined;

$$\textit{Distance measure} \qquad \varepsilon^2 = \left\|\Phi - \Phi_f\right\|^2 \tag{4.24}$$

is the distance between the mean subtracted image and the reconstructed image.

The recognition of an image knowing these two measures δ_i and ε can be done as follows:

(a) If $\varepsilon > \Theta \Rightarrow$ the image is not a face (independent of the values of δ_i),

(b) If $\varepsilon < \Theta$ and for all i $\delta_i > \Theta \Rightarrow$ the image is an unknown face,

(c) If $\varepsilon < \Theta$ and for one of i $\delta_i < \Theta \Rightarrow$ the image belongs to the training face class i.

This is the standard Eigenface approach suggested by Turk and Pentland in 1991 and variations of this method is explained in the next sections.

4.2. Pure LDA Method

The second method implemented throughout the study is the pure LDA [44] method, where Fisher's Linear Discrimant is applied to classify the face image vectors. Implementing LDA to the original data just *discriminate*s the data. Hence, it is an alternative to the Eigenface approach where the data was *generalize*d.

There are various measures available for quantifying the discriminatory power. A commonly used one is the ratio of the between-class scatter matrix of the projected data to the within-class scatter matrix of the projected data;

Discriminatory Power
$$J(T) = \frac{\left| T^T \cdot S_b \cdot T \right|}{\left| T^T \cdot S_w \cdot T \right|} \tag{4.25}$$

where S_b is the between-class and S_w is the within-class scatter matrix.

The image is represented by (N_x x N_y) pixels.

Image
$$I: \quad (N_x \text{ x } N_y) \text{ pixels} \tag{4.26}$$

As in the case of PCA, the image matrix I of size (N_x x N_y) pixels is converted to the image vector Γ of size (P x 1) where P = (N_x x N_y); that is the image matrix is reconstructed by adding each column one after the other.

Training Set
$$\Gamma = \left[\Gamma_1 \quad \Gamma_2 \quad \ldots \quad \Gamma_{Mt}\right] \tag{4.27}$$

is the training set of image vectors and its size is (P x M_t) where M_t is the number of the training images.

Linear Discriminant Analysis can not deal with the problem of one training image per class case. In that case, S_w turns out to be an identity matrix; hence the solution reduces to a standard Eigenface approach with S_w being the covariance matrix [21].

In LDA, mean face images are also calculated for each face class; this is due to need for the calculation of each face classes inner variation. Hence, for each of c individuals having q_i training images in the database

Class Mean Face
$$\Psi_{Ci} = \frac{1}{q_i} \sum_{k=1}^{q_i} \Gamma_k \tag{4.28}$$

is the arithmetic average of the training image vectors corresponding to the same individual at each pixel point for each class; i = 1, 2, …, c and its size is (P x 1).

Moreover, mean face is calculated from the arithmetic average of all the training image vectors at each pixel point;

Mean Face $$\Psi = \frac{1}{M_t}\sum_{k=1}^{M_t}\Gamma_k \tag{4.29}$$

and its size is (P x 1).

Mean subtracted image $$\Phi_i = \Gamma_i - \Psi_{Ci} \tag{4.30}$$

is the difference of the training image from its class mean face and its size is (P x 1).

For c individuals having q_i training images, the within-class scatter matrix is computed as;

Within Class Scatter Matrix $$S_w = \sum_{i=1}^{c} P(C_i)\Sigma_i \tag{4.31}$$

which represents the average scatter Σ_i of the image vectors of different individuals C_i around their respective class means Ψ_{Ci}. The size of S_w is (P x P).

$P(C_i)$ is the prior class probability and may be written as

Prior class probability $$P(C_i) = \frac{1}{c} \tag{4.32}$$

with the assumption that each class has equal prior probabilities.

Average Scatter $$\Sigma_i = E\left[\Phi_i \cdot \Phi_i^T\right] = E\left[(\Gamma_i - \Psi_{Ci}) \cdot (\Gamma_i - \Psi_{Ci})^T\right] \tag{4.33}$$

Similarly, between-class scatter matrix is computed as;

Between-Class Scatter Matrix $$S_b = \sum_{i=1}^{c} P(C_i)(\Psi_{Ci} - \Psi) \cdot (\Psi_{Ci} - \Psi)^T \tag{4.34}$$

which represents the scatter of each class mean Ψ_{Ci} around the overall mean vector Ψ, and its size is (P x P).

The objective is to maximize *J(T)*; in other words finding an optimal projection *W* which maximizes between-class scatter and minimizes within-class scatter.

$$W = \arg\max_{T}\left(J(T)\right) \Rightarrow \max\left(J(T)\right) = \frac{\left|T^T \cdot S_b \cdot T\right|}{\left|T^T \cdot S_w \cdot T\right|}\Big|_{T=W} \tag{4.35}$$

W can be obtained by solving the generalized eigenvalue problem [44];

$$S_b W = S_w W \lambda_w \tag{4.36}$$

From the generalized eigenvalue equation, only *c-1* or less of the eigenvalues come out to be nonzero. This is due to the fact that S_w is the sum of *c* matrices of rank one or less, and because at most *c-1* of these are linearly independent. As a result, no more than *c-1* of the eigenvalues are nonzero, and only the eigenvectors coming out with these nonzero eigenvalues can be used in forming the *W* matrix and the size of the *W* matrix is (P x *(c-1)*).

Next, the training image vectors are projected to the classification space by the dot product of the optimum projection *W* and the image vector as follows;

Classification space projection $g\left(\Phi_i\right) = W^T \cdot \Phi_i$ (4.37)

is the projection of the training image vectors to the classification space which is of size (*(c-1)* x 1) where i = 1, 2, …, M_t

Next, the eigenface projection of the test image vector is projected to the classification space in the same manner;

Classification space projection $g(\Phi_T) = W^T \cdot \Phi_T$ (4.38)

which is of size (*(c-1)* x 1).

Finally, the distance between the projections is calculated by the Euclidean distance between the training and test classification space projections;

Distance measure $d_{Ti} = \|g(\Phi_T) - g(\Phi_i)\| = \sqrt{\sum_{k=1}^{c} (g_k(\Phi_T) - g_k(\Phi_i))^2}$ (4.39)

is the distance measure which is scalar and calculated for i = 1, 2, …, M_t

The test image is assumed to be in the class whose distance is the minimal among all other class distances.

4.3. Subspace LDA Method

The next method implemented in this study is the Subspace LDA method [21], [22], [23], which is simply the implementation of PCA by projecting the data onto the eigenface space and then implementing LDA to classify the eigenface space projected data. Projecting the data to the eigenface space *generalize*s the data, whereas implementing LDA by projecting the data to the classification space *discriminate*s the data. Thus, Subspace LDA approach seems to be a complementary approach to the Eigenface method.

PCA and LDA were described in the previous sections, hence here just the implementation shall be described.

The image is represented by (N_x x N_y) pixels.

Image I : (N_x x N_y) pixels (4.40)

The image matrix I of size (N_x x N_y) pixels is converted to the image vector Γ of size (P x 1) where P = (N_x x N_y);

Training Set $$\Gamma = \begin{bmatrix} \Gamma_1 & \Gamma_2 & \dots & \Gamma_{Mt} \end{bmatrix} \quad (4.41)$$

is the training set of image vectors and its size is (P x M_t) where M_t is the number of the training images.

Mean Face $$\Psi = \frac{1}{M_t} \sum_{i=1}^{M_t} \Gamma_i \quad (4.42)$$

is the arithmetic average of the training image vectors at each pixel point and its size is (P x 1)

Mean subtracted image $$\Phi = \Gamma - \Psi \quad (4.43)$$

is the difference of the training image from the mean image (size P x 1)

Difference Matrix $$A = \begin{bmatrix} \Phi_1 & \Phi_2 & \dots & \Phi_{Mt} \end{bmatrix} \quad (4.44)$$

is the matrix of all the mean subtracted training image vectors and its size is (P x M_t).

Covariance Matrix $$X = A \cdot A^T = \frac{1}{M_t} \sum_{i=1}^{M_t} \Gamma_i \Gamma_i^T \quad (4.45)$$

is the covariance matrix of the training image vectors of size (P x P).

Assume, the eigenvectors υ_i of the covariance matrix are obtained using the equation 4.13 in Section 4.1. The training image vectors can be projected to the eigenface space and thus the weight of each eigenvector to represent the image in the eigenface space is calculated;

Projection $$\omega_k = \upsilon_k^T \cdot \Phi = \upsilon_k^T \cdot (\Gamma - \Psi) \quad (4.46)$$

is the projection of a training image on each of the eigenvectors where k = 1, 2, …, M′ (using M′ is an alternative for M_t was described in Section 4.1).

Weight Matrix $$\Omega = [\omega_1 \quad \omega_2 \quad ... \quad \omega_{M'}]^T \tag{4.47}$$

is the representation of the training image in the eigenface space and its size is (M′ x 1).

By performing all these calculations, the training images are projected onto the eigenface space; that is a transformation from P-dimensional space to M′ dimensional space. This PCA step is achieved to reduce the dimension of the data, also may be referred as a feature extraction step. From this step on, the each image is an (M′ x 1) dimensional vector in the eigenface space.

With the projected data in hand, a new transformation is performed; the classification space projection by LDA. Instead of using the pixel values of the images (as done in pure LDA), the eigenface projections are used in the Subspace LDA method.

Again, as in the case of pure LDA, a discriminatory power is defined as;

Discriminatory Power $$J(T) = \frac{\left|T^T \cdot S_b \cdot T\right|}{\left|T^T \cdot S_w \cdot T\right|} \tag{4.48}$$

where S_b is the between-class and S_w is the within-class scatter matrix.

For c individuals having q_i training images in the database, the within-class scatter matrix is computed as;

Within Class Scatter Matrix $$S_w = \sum_{i=1}^{c} P(C_i)\Sigma_i \tag{4.49}$$

which represents the average scatter Σ_i of the projection Ω in the eigenface space of different individuals C_i around their respective means m_i. The size of S_w depends on the size of the eigenface space; if M′ of the eigenfaces were used, then the size of S_w is (M′ x M′).

Here, eigenface space class mean is defined as;

Eigenface Class Mean
$$m_i = \frac{1}{q_i}\sum_{k=1}^{q_i}\Omega_k \tag{4.50}$$

is the arithmetic average of the eigenface projected training image vectors corresponding to the same individual; $i = 1, 2, \ldots, c$ and its size is (M′ x 1).

Moreover, mean face is calculated from the arithmetic average of all the projected training image vectors;

Eigenface Mean Face
$$m_0 = \frac{1}{M_t}\sum_{k=1}^{M_t}\Omega_k \tag{4.51}$$

The average scatter is calculated as;

Average Scatter
$$\Sigma_i = E\left[(\Omega - m_i)\cdot(\Omega - m_i)^T\right] \tag{4.52}$$

Also between-class scatter matrix is computed as;

Between-Class Scatter Matrix S_b
$$= \sum_{i=1}^{c} P(C_i)(m_i - m_0)\cdot(m_i - m_0)^T \tag{4.53}$$

which represents the scatter of each projection classes mean m_i around the overall mean vector m_0 and its size is (M′ x M′).

$P(C_i)$ is the prior class probability and may be written as

Prior class probability
$$P(C_i) = \frac{1}{c} \tag{4.54}$$

with the assumption that each class has equal prior probabilities.

The objective is to maximize *J(T)*; that is, to find an optimal projection *W* which maximizes between-class scatter and minimizes within-class scatter.

$$W = \arg\max_{T}\left(J(T)\right) \Rightarrow \max\left(J(T)\right) = \frac{\left|T^T \cdot S_b \cdot T\right|}{\left|T^T \cdot S_w \cdot T\right|}\Big|_{T=W} \tag{4.55}$$

Then, *W* can be obtained by solving the generalized eigenvalue problem;

$$S_b W = S_w W \lambda_w \tag{4.56}$$

Next, the eigenface projections of the training image vectors are projected to the classification space by the dot product of optimum projection *W* and weight vector.

Classification space projection $g(\Omega_i) = W^T \cdot \Omega_i$ (4.57)

is the projection of the training image vectors' eigenface projections to the classification space which is of size (*(c-1)* x 1) where i = 1, 2, …, M_t

The training stage is completed in this step. Next, the test image is taken;

Test image vector Γ_T (4.58)

is the test image vector of size (P x 1).

Mean subtracted image $\Phi_T = \Gamma_T - \Psi$ (4.59)

is the difference of the test image from the mean image (size P x 1).

Projection $\omega_k = \upsilon_k^T \cdot \Phi_T = \upsilon_k^T \cdot (\Gamma_T - \Psi)$ (4.60)

is the projection of a training image on each of the eigenvectors where k = 1, 2, …, M′

Weight Matrix $$\Omega_T = \begin{bmatrix} \omega_1 & \omega_2 & \dots & \omega_{M'} \end{bmatrix}^T \quad (4.61)$$

is the representation of the test image in the eigenface space and its size is (M′ x 1).

Then, the eigenface projection of the test image vector (i.e. weight matrix) is projected to the classification space in the same manner.

Classification space projection $$g(\Omega_T) = W^T \cdot \Omega_T \quad (4.62)$$

which is of size (*(c-1)* x 1).

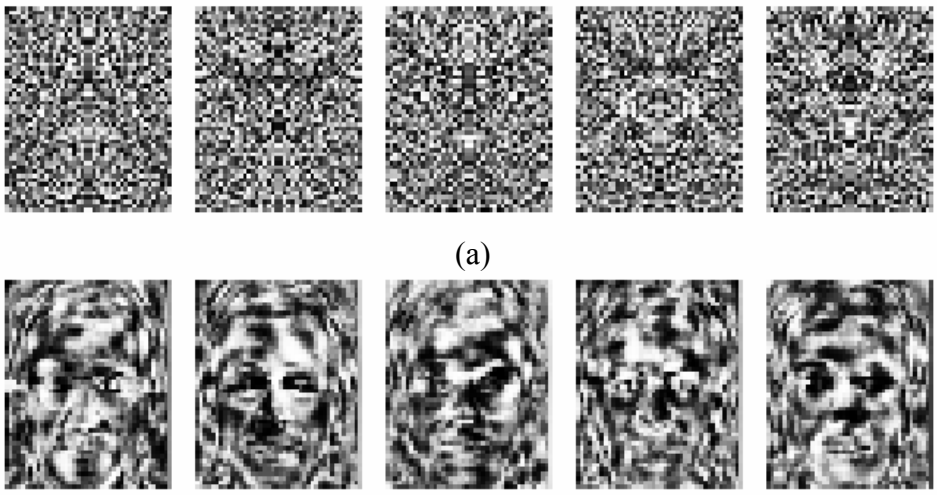
Finally, the distance between the projections is determined by the Euclidean distance between the training and test classification space projections;

Distance measure $$d_{Ti} = \left\| g(\Omega_T) - g(\Omega_i) \right\| = \sqrt{\sum_{k=1}^{c} \left(g_k(\Omega_T) - g_k(\Omega_i) \right)^2} \quad (4.63)$$

is the distance measure which is scalar and calculated for i = 1, 2, …, M_t

Some other distance measures (strictly weighted and softly weighted) are also suggested in [21], but that measures were not studied in this thesis.

Example LDA and Subspace LDA bases are given in Figure 4.5.



(a)

(b)

Figure 4.5: (a) Pure LDA bases, **(b)** Subspace LDA bases [21].

4.4. Bayesian PCA Method

The last method implemented in this study is the Bayesian PCA method [24], [25], [26], [27], [28], [29], [30], [31], which is simply implementing PCA by projecting the data onto the eigenface space and then utilizing a Bayesian classifier to classify the eigenface space projected data. Applying PCA for dimensionality reduction condenses the original image space into a compact one with the merits of preventing noises and enhancing generalization as well as solving the curse of dimensionality problem. Then, the Bayes Classifier, which yields the minimum error when the underlying probability density functions are known, carries out the classification in the eigenface space using the Maximum A Posteriori (MAP) or Maximum Likelihood (ML) rules.

The image is represented by (N_x x N_y) pixels.

Image $I : \quad (N_x \text{ x } N_y)$ pixels (4.64)

The image matrix I of size (N_x x N_y) pixels is converted to the image vector Γ of size (P x 1) where P = (N_x x N_y); that is the image matrix is reconstructed by adding each column one after the other.

Training Set $$\Gamma = \left[\Gamma_1 \quad \Gamma_2 \quad \ldots \quad \Gamma_{Mt}\right] \tag{4.65}$$

is the training set of image vectors and its size is (P x M_t) where M_t is the number of the training images.

In this method, the data used is the face image differences (not the face images themselves as in the PCA and Subspace LDA methods), denoted by

Image difference $$\Delta = I_1 - I_2 \tag{4.66}$$

which is characteristic of typical variations in the appearance of the same face image. It can be stated here that, there should be at least two training face images in the database for obtaining an image difference.

Difference Mean $$\Psi_\Delta = \frac{2}{M_t} \sum_{i=1}^{M_t/2} \Delta_i \tag{4.67}$$

is the arithmetic average of the training image vector differences at each pixel point and its size is (P x 1). In this case, it is assumed that there are 2 images per individual in order not to make the problem more complex.

Mean Normalized Difference $$\widetilde{\Delta} = \Delta - \Psi_\Delta \tag{4.68}$$

is the mean normalized difference needed for the calculations, and its size is (P x 1).

Two classes of facial image variations are defined;

Intrapersonal variations Ω_I (4.69)

which corresponds to variations of the face images of the same individual (e.g. facial expression, lighting, etc.) and

Extrapersonal variations Ω_E (4.70)

which corresponds to variations of the face images of different individuals.

It is assumed that both variations are Gaussian distributed [32]. Hence, the estimates of the likelihood functions $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ for a given intensity difference $\Delta = I_1 - I_2$ should be obtained.

The similarity measure for retrieval is represented in terms of the a posteriori probability;

Similarity $$S(I_1, I_2) = P(\Delta \in \Omega_I) = P(\Omega_I|\Delta) \tag{4.71}$$

where $P(\Omega_I|\Delta)$ is the intrapersonal a posteriori probability given by Bayes rule;

$$S(I_1, I_2) = P(\Omega_I|\Delta) = \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)} \tag{4.72}$$

where $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ are the estimates of the likelihoods, which should be derived from the training face images using PCA.

The prior probability $P(\Omega_I)$ is set to the ratio of the number of training images of an individual vs. the total number of training images, and $P(\Omega_E)$ is set to the remaining training images vs. the total number of training images in the database; as $P(\Omega_I)$ and $P(\Omega_E)$ should add up to unity.

From the above similarity equation, it can be stated that the standard face recognition problem has turned into a binary pattern

classification problem with Ω_I and Ω_E. This problem is solved using two alternative methods;

1) The maximum a posteriori (MAP) rule: Two images belong to the same individual if

$$P(\Omega_I|\Delta) > P(\Omega_E|\Delta) \tag{4.73}$$

or from another point of view;

$$S(I_1, I_2) > 0.5 \tag{4.74}$$

2) The maximum likelihood (ML) estimation: Only intrapersonal likelihood is used for the similarity measure, which can be written as;

$$S'(I_1, I_2) = P(\Delta|\Omega_I) \tag{4.75}$$

where the test image is compared to each training image in the database and classified in the class where S' is the biggest among all.

In this method, the likelihoods $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ should be calculated and they are obtained from the training face images using PCA. The derivation is explained in Section 4.4.1.

4.4.1. Estimation of the Likelihood of a Face

First, assuming the differences of face image vectors having Gaussian distributions, the likelihood of an input face image difference can be written as [32];

$$P(\Delta|\Omega) = \frac{e^{\left(-\frac{1}{2}(\Delta-\Psi_\Delta)^T \Sigma^{-1}(\Delta-\Psi_\Delta)\right)}}{(2\pi)^{N/2}|\Sigma|^{1/2}} \tag{4.76}$$

where Σ is the covariance matrix.

The Mahalanobis distance from the above equation is;

$$\textit{Mahalanobis distance } d(\Delta) = \tilde{\Delta}^T \Sigma^{-1} \tilde{\Delta} = (\Delta - \Psi_\Delta)^T \Sigma^{-1} (\Delta - \Psi_\Delta) \qquad (4.77)$$

which is the measure of the separation of a face image vector from the face data set in the image space.

However, instead of calculating this likelihood in the image space; it can be calculated in the eigenface space. The mathematical formulation behind is as follows: The eigenvalue equation in the standard PCA method was;

$$\mathrm{Y} \cdot \nu_i = \mu_i \cdot \nu_i \qquad (4.78)$$

it can be rewritten in the form of

$$\Sigma \cdot V = V \cdot \Lambda \qquad (4.79)$$

where Σ is the covariance matrix (called Y in the standard PCA case), V is the eigenvector matrix of Σ, and Λ is the corresponding diagonal matrix of eigenvalues of the equation. The unitary matrix V defines a coordinate transformation from the image space to the eigenface space;

$$\Omega = V^T \cdot \tilde{\Delta} \qquad (4.80)$$

As V is unitary;

$$V^T \cdot V = V \cdot V^T = I \qquad (4.81)$$

$$\Sigma \cdot V = V \cdot \Lambda \qquad (4.82)$$

$$V^T \cdot \Sigma \cdot V = V^T \cdot V \cdot \Lambda \qquad (4.83)$$

$$\Lambda = V^T \cdot \Sigma \cdot V \qquad (4.84)$$

This equation tells about the structure of the eigenvalue equation; a covariance matrix is left multiplied by the transpose of the eigenvector matrix and right multiplied by the eigenvector matrix itself; and the resulting equation gives the eigenvalues.

This equation shall be used by a few arrangements in the Mahallanobis distance;

Mahalanobis distance $$d(\Delta) = \widetilde{\Delta}^T \Sigma^{-1} \widetilde{\Delta} \tag{4.85}$$

$$d(\Delta) = \widetilde{\Delta}^T \left[V \cdot \Sigma^{-1} \cdot V^T \right] \widetilde{\Delta} \tag{4.86}$$

$$d(\Delta) = \widetilde{\Delta}^T \left[V \cdot \Sigma^{-1} \cdot V^T \right] \widetilde{\Delta} \tag{4.87}$$

$$d(\Delta) = \Omega^{-1} \Lambda^{-1} \Omega \tag{4.88}$$

as $\Omega = V^T \cdot \widetilde{\Delta}$.

Due to the diagonalized form; the final equation can be expressed as;

$$d(\Delta) = \sum_{i=1}^{N} \frac{\omega_i^2}{\lambda_i} \tag{4.89}$$

Although this distance seems to be simple; it is still computationally very complex due to high dimension of the image space, N. Hence, this measure is decomposed into two parts by dividing the space into two orthogonal spaces; the principal eigenface space F (made up of M′ eigenvectors), and its orthogonal space $\overline{F}$ (made up of the remaining eigenvectors, M′+1 to N, which were not used in ordinary PCA).

$$d(\Delta) = \sum_{i=1}^{M'} \frac{\omega_i^2}{\lambda_i} + \sum_{i=M'+1}^{N} \frac{\omega_i^2}{\lambda_i} \tag{4.90}$$

The first term of this distance measure can be computed by the projections of the data in the eigenface space constructed using the first M′ eigenvectors. However, the second term is computationally difficult to compute. In fact, as there is not enough data to construct the covariance matrix Σ, the eigenvalues λ_i corresponding to the eigenvectors larger than M are zero.

Hence, another approach to solve this ambiguity is required. The values of λ_i are unknown but their sum can be found from the residual reconstruction error of the eigenface projection;

$$\textit{Residual reconstruction error } \varepsilon^2(\Delta) = \sum_{i=M'+1}^{N} \omega_i^2 = \left\|\widetilde{\Delta}\right\|^2 - \sum_{i=1}^{M'} \omega_i^2 \tag{4.91}$$

as

$$\left\|\widetilde{\Delta}\right\|^2 = \|\Omega\|^2 = \sum_{i=1}^{M'} \omega_i^2 + \sum_{i=M'+1}^{N} \omega_i^2 \tag{4.92}$$

assuming the image is exactly projected in the eigenface space.

Here, an approximation is needed for the sum of λ_i. Thus, an estimator for the Mahalanobis distance may be written as;

$$\hat{d}(\Delta) = \sum_{i=1}^{M'} \frac{\omega_i^2}{\lambda_i} + \frac{1}{\rho}\left[\sum_{i=M'+1}^{N} \omega_i^2\right] = \sum_{i=1}^{M'} \frac{\omega_i^2}{\lambda_i} + \frac{\varepsilon^2(\Delta)}{\rho} \tag{4.93}$$

where

$$\rho = \frac{1}{N - M'} \sum_{i=M'+1}^{N} \lambda_i \tag{4.94}$$

Finally, the likelihood term (in fact, the estimate of the likelihood term) may be written according to the Mahalanobis distance;

$$\hat{P}(\Delta|\Omega)=\left[\frac{e^{\left(-\frac{1}{2}\sum_{i=1}^{M'}\frac{\omega_i^2}{\lambda_i}\right)}}{(2\pi)^{M'/2}\prod_{i=1}^{M'}\lambda_i^{1/2}}\right]\cdot\left[\frac{e^{\left(-\frac{\varepsilon^2(\Delta)}{2\rho}\right)}}{(2\pi\rho)^{(N-M')/2}}\right]=P_F(\Delta|\Omega)\hat{P}_{\overline{F}}(\Delta|\Omega) \qquad (4.95)$$

where $P_F(\Delta|\Omega)$ is the true marginal density in F space, and $\hat{P}_{\overline{F}}(\Delta|\Omega)$ is the estimated marginal density in the $\overline{F}$ space.

Finally, $\hat{P}(\Delta|\Omega_I)$ equation above is used to obtain the likelihoods $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$, and these likelihoods are used in classifying the images in a MAP or ML manner.

CHAPTER 5

SIMULATIONS

In this chapter, PCA [10], LDA [44], Subpace LDA [21], [22], [23], and Bayesian PCA [24], [25], [26], [27], [28], [29], [30], [31] algorithms are tested individually, and their performances are compared relative to changes in facial expression, aging, lighting and the effects of using different camera on different databases.

The utilized face databases are ORL [55] and FERET [56], [57], [58], and a test database constructed from the speakers of the CNN-TURK TV channel.

The tests are conducted via writing programs in Matlab. Another choice could be to write codes in C/C++. However, the easiness of Matlab for mathematical applications made Matlab the choice.

5.1. Utilized Face Databases

5.1.1. Olivetti Research Laboratory (ORL) Face Database

The ORL face database [55] is made up of a set of faces taken between April 1992 and April 1994 at the Olivetti Research Laboratory in Cambridge, UK.

There are 10 different images of 40 distinct subjects. For some of the subjects, the images were taken at different times, varying lighting slightly, facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). All the images are taken against a dark homogeneous background and the subjects are in up-right, frontal position (with tolerance for some side movement).

The files are in .TIF file format (Tagged Image File Format). The size of each image is 92 x 112 (width x height), 8-bit grey levels. The images are renamed as;

0x.TIF

where x ranges from 568 to 967.

As there are 10 images per individual in the ORL database, first 5 of them are used to train the systems and the rest are used to test them. Moreover, different tests are conducted, such as changing the training and test images used for each individual and using less or more number of training and test images in order to test the systems reaction to these changes.

5.1.2. The Face Recognition Technology (FERET) Face Database

The FERET database [56], [57], [58], is a standard database that is used for algorithm development and testing. The FERET program is sponsored by the Department of Defense Counterdrug Technology Development program through the Defense Advanced Research Projects Agency (DARPA), with the U.S. Army Research Laboratory (ARL) serving as technical agent.

The images were collected in 15 sessions between August 1993 and July 1996. Collection sessions lasted one or two days. In order to

maintain consistency throughout the collection of database, the same physical setup and location was used in each photography session. However, as the equipment had to be reassambled in each session, there were variations from session to session.

Images of an individual were acquired in sets of 5-11 images. Two frontal images were taken (fa and fb), where fb is an alternative facial expression of fa. For 200 sets of images, a third frontal image was taken with a different camera and different lighting (fc). The remaining images were taken at various aspects between right and left profile: right and left profile (labeled pr and pl), right and left quarter profile (qr, ql), and right and left half profile (hr, hl). Additionally, five extra locations (ra, rb, rc, rd, and re), irregularly spaced among the basic images, were collected. In order to add variations to the database, photographers sometimes took a second set of images for which the subjects were asked to put on their glasses and/or pull their hair back. Moreover, a different set of images of a person was taken on a later date (duplicate). These images result in variations in scale, pose, expression, and illumination of the face. Figure 5.1 shows different camera angles used for a subject face.

The FERET database released in March 2001, consists of 14051 eight-bit greyscale images from 1196 individuals.

The images are stored in .TIF format and as raw 8-bit data. They are 256 x 384 (width x height). Attempts were made to keep the interocular distance (the distance between the eyes) of each subject to between 40 and 60 pixels. The images consist primarily of an individuals head, neck, and sometimes the upper part of the shoulders.

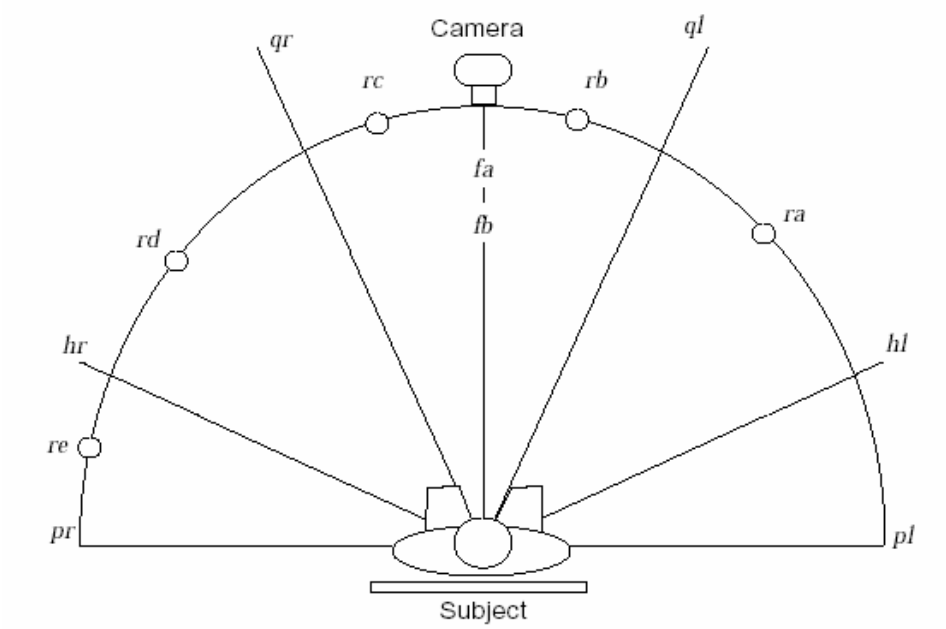


Figure 5.1: Possible camera angles collected for a subject face [57].

The images are collected in a semi-controlled environment. In order to maintain a degree of consistency throughout the database, the same physical setup is used in each photography session. However, there is some variation over collections from site to site, since the equipment must be reassembled for each session.

The naming convention for the FERET imagery in the distribution is of the form nnnnnxxfffq_yymmdd.ext where:

1. nnnnn is a five digit integer that uniquely identifies the subject
2. xx is a two lowercase character string that indicates the kind of imagery, where
   - fa indicates a regular frontal image

- fb indicates an alternative frontal image, taken seconds after the corresponding fa
- ba is a frontal images which is entirely analogous to the fa series
- bj is an alternative frontal image, corresponding to a ba image, and analogous to the fb image
- bk is also a frontal image corresponding to ba, but taken under different lighting
- bb through bi is a series of images taken with the express intention of investigating pose angle effects. Specfically, bf - bi are symmetric analogues of bb - be.
- ra through re are “random” orientations. Their precise angle is unknown. It appears that the pose angles are random but consistent. The pose angles in the table were derived by manual measurement of inter-eye distances in the image, and in their corresponding frontal image.

3. fff is a set of three binary (zero or one) single character flags. In order these denote:
   - First flag indicates indicates whether the image is releasable for publication. The flag has fallen into disuse: All images are available via this CD-ROM distribution, but still none may be published without the explicit written permission of the government.
   - Second flag indicates that the image is histogram adjusted if this flag is 1

- Third flag indicates whether the image was captured using ASA 200 or 400 film, 0 implies 200.

4. q is a modifier that is not always present. When it is, the meanings are as follows:

   - (a) Glass worn. This flag is a sufficient condition only, the images of subjects wearing glasses do not necessarily carry this flag. Some retroactive re-truthing of such images to fix this problem is warranted.
   - (b) Duplicate with different hair length.
   - (c) Glasses worn and different hair length
   - (d) Electronically scaled (resized) and histogram adjusted.
   - (e) Clothing has been electronically retouched.
   - (f) Image brightness has been reduced by 40%
   - (g) Image brightness has been reduced by 80%
   - (h) Imagesize has been reduced by 10%, with white border replacement
   - (i) Image size has been reduced by 20%, with white border replacement
   - (j) Image size has been reduced by 30%, with white border replacement

   Note that the modifications d through j are the result of applying various off-line operations to real images in the database; the "parent" image is that image without the "q" modifier present at all.

5. The three fields are the date that the picture was taken in year, month, and day format.
6. The filename extension is .tif. The images on the CDROMs carry an additional .bz2 suffix that indicates that the files have been losslessly compressed using the bzip2 compressor.

Figure 5.2: Example frontal images from the FERET database corresponding to one individual.

The FERET tests can be conducted using the gallery (training) and probe (test) images suggested by the FERET team in the FERET CD-ROM. Table 5.1 shows the suggested gallery and probe images in the database and which probe names test which evaluation task. The gallery set is the same for all the four test setups. Such a suggestion is stated in order to standardize the test results.

It is important to point out here that the gallery and probe sets in Table 5.1 use only the frontal images which is also the main concern of this theses (Figure 5.2). Also note that the tests used a single gallery

containing 1196 images. The Duplicate I probe images (dup1) were obtained anywhere between one minute and 1031 days (34 months) after their respective gallery matches. The harder Duplicate II probe images (dup2) are a strict subset of the Duplicate I images; they are those taken between 540 and 1031 days (18 months) after their gallery entries. There is usually only a few seconds between the capture of the gallery-probe pairs in facial expression evaluation probe images (fafb).

Evaluation Task	Recognized Names	Gallery (1196)	Probe Set
Aging of subjects	Duplicate I	gallery.names	probe_dup_1_*.names (722)
Aging of subjects	Duplicate II	gallery.names	probe_dup_2_*.names (234)
Facial Expression	fafb	gallery.names	probe_fafb_*.names (1195)
Different camera and Illumination	fafc	gallery.names	probe_fafc_*.names (194)

Table 5.1: The Gallery and Probe Sets used in the standard FERET test in September 1996.

5.1.3. CNN-TURK Speakers Face Database

The CNN-TURK Speakers Face Database is collected during this thesis study, in order to test the systems performances in a real-life application. The images are first captured from moving scenes of a TV, and then converted to 8-bit grey scale .BMP images. Finally, they were manualy cropped in order the image to contain the face region.

The naming convention for the CNN-TURK Speakers Face Database is of the form CNNa_b.BMP where:

1. a is 1 or 2 digit number ranging from 1-11 representing each individual.
2. b has 1 or 2 digits where images taken in the first week range from “1” to “18” and the images (taken 7-10 days later) range from “t1” to “t5”.

Figure 5.3: Training and gallery sets of two individuals from the CNN -TURK Speakers face database.

As there are 11 individuals in this database each having 23 frontal face images, it can be suggested to use the first 5 images as the training set and the other images taken on the same day (i.e. 6-18) as the test set for the expression test, and t1-t15 as the test set for the illumination change and aging tests (shown in Figure 5.3).

As these images are taken from TV, they help us test facial expression change, illumination change. Moreover, due to interlacing, smearing or somehow visually distorted frames are observed in this database.

5.2. Preprocessing Techniques

In this thesis, mainly the effect of utilizing 5 preprocessing steps is examined: Resizing, rotation correction, cropping, histogram equalization, and masking.

The position of eye, nose and mouth are given for the FERET database images, so these preprocessing techniques are applied on FERET images automatically. The preprocessing techniques are applied in the FERET database in the order of rotation correction, cropping, histogram equalization, and masking. Some example images from the FERET database are given in Figure 5.4, in order to show the effect of each preprocessing step.

Resizing and histogram equalization are applied on ORL and CNN-TURK Speakers face databases automatically, and cropping and rotation correction are applied on CNN-TURK Speakers face database manually.

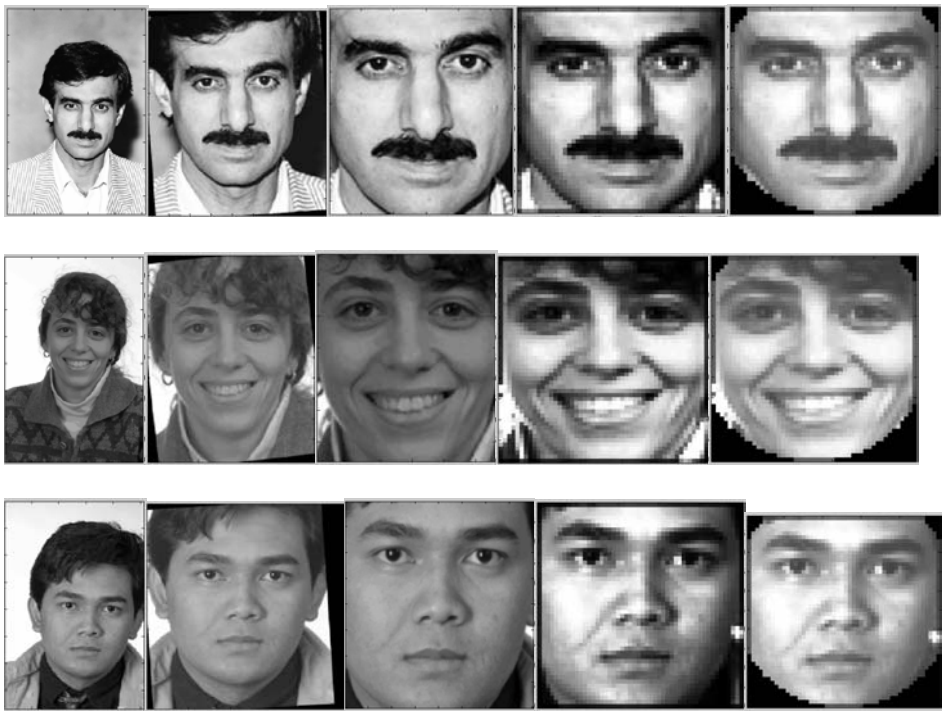


Figure 5.4: The effect of the preprocessing steps on FERET images. (From left to right; original, rotated, cropped, histogram equalized and masked image).

5.2.1. Rotation Correction

When the locations of eyes, nose and mouth are given for a face image, this data can be used to rotate the image so that all the face images are exactly positioned the same.

When the positions for the right and left eyes are known, the inverse tangent of the angle between the lines, l_1 and l_2, connecting the mid-points of two eyes can be calculated. The image can be rotated using the calculated angle. This process is drawn in Figure 5.5. l_1 is the initial line connecting the mid-points of the eyes, and l_2 is the final line connecting the mid-points of the eyes. The locations of eyes, nose and

mouth are not given for the ORL database; hence, a rotation is not performed (even manually). The rotation correction is done manually in the CNN-TURK Speakers Face Database. On the other hand, a rotation correction is needed in the FERET database where the eye, nose and mouth locations are given in the distributed CD-ROM of the database. Hence, FERET face images are automatically rotated using the described procedure.

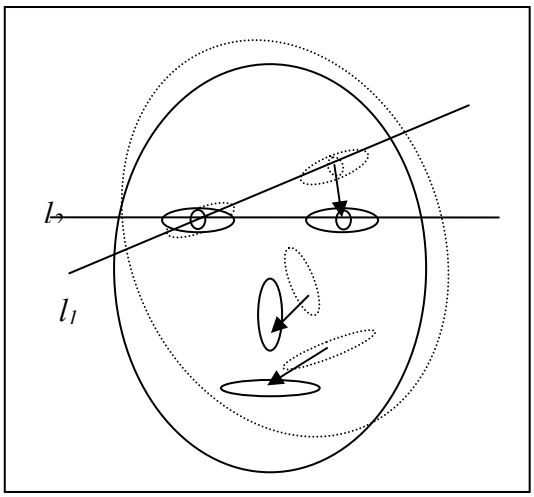


Figure 5.5: The rotation correction procedure when eye, nose and mouth locations are given.

5.2.2. Cropping

When the area of an image is much larger compared to that of a face, the region of the image where the face is located is cut out from the image and only this area is used in the process of face recognition. In this study, the face area is determined as (given in Figure 5.6);

- Left and right borders are determined by using half of the distance between the right and left eyes horizontal position. This distance

is added to right eyes horizontal position and subtracted from the left eyes horizontal position.

- Top and bottom borders are determined using the half of the distance between the left eyes and mouths vertical position. This distance is added to left eyes vertical position and subtracted from the left eyes vertical position.

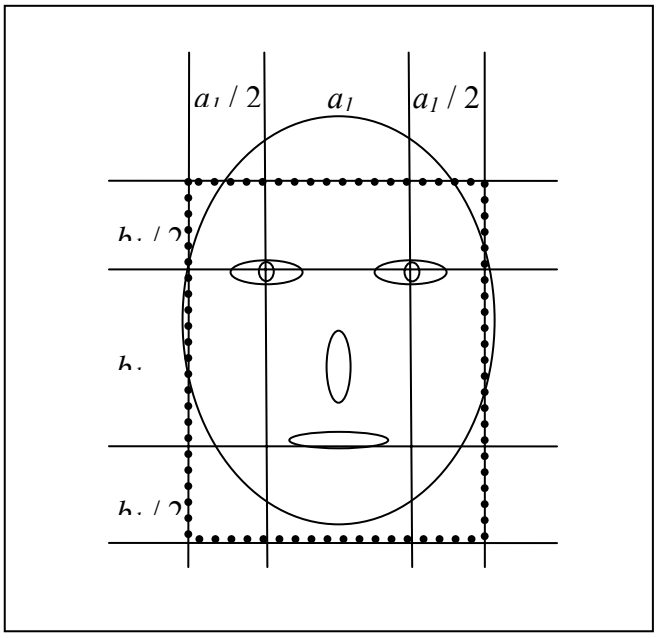


Figure 5.6: The cropping procedure when eye, nose and mouth locations are given.

Again, cropping is not applied on ORL database and manually applied on CNN-TURK Speakers database. On the other hand, cropping is applied automatically (as in the case of rotation) to the face images in the FERET database and the results are observed.

5.2.3. Histogram Equalization

Histogram equalization is applied in order to improve the contrast of the images. The peaks in the image histogram, indicating the commonly used grey levels, are widened, while the valleys are compressed.

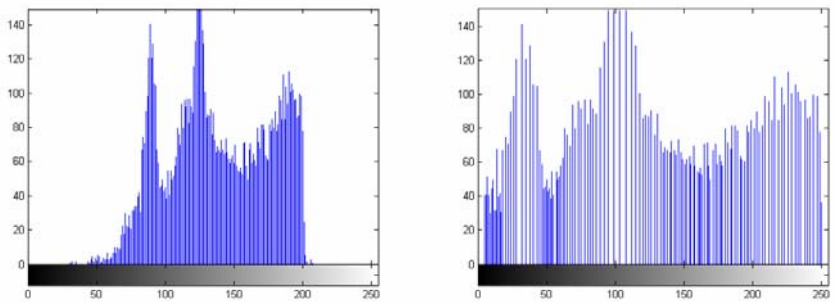


Figure 5.7: The histogram of an image before (left) and after (right) the histogram equalization.

Two histogram plots are given in Figure 5.7. The histogram on the left is before histogram equalization (between 6-250) is applied and the one on the right is after histogram equalization is applied. Histogram equalization is applied to all the three databases automatically.

5.2.4. Masking

By using a mask, which simply has a face shaped region, the effect of background change is minimized. The effect of masking is studied only on FERET database images. The mask used in this study is shown in Figure 5.8.

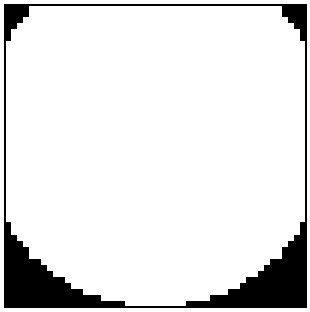


Figure 5.8: The face shaped mask used throughout the study.

5.3. Test Results

5.3.1. Tests Performed Using the ORL Face Database

5.3.1.1. Using Different Number of Training and Test Images

Since there are 10 images per individual in the ORL database the effect of different number of training and test image combinations is tested. The tests are performed using PCA method [10].

Total Number of Images	Number of Training Images (per individual)	Number of Test Images (per individual)	Success Rate (%)
400	1	9	69,7
400	5	5	88,0
400	9	1	87,5

Table 5.2: The success rates using different number of training and test images.

From the tests results in Table 5.2, it can be argued that using 5 training images and 5 testing images for each individual result with

meaningful results; hence, the rest of the ORL tests are conducted using 5 training and 5 test images per individual.

5.3.1.2. Changing the Training and Test Image Combination

In the second part of the experiments, the effect of changing the images of each individual used in the training and testing stage in a rotating manner has been studied. This stage of the experiments is conducted using the Subspace LDA.

From the tests results in Table 5.3, it can be easily observed that the success rate changes with respect to the utilized sets of training and testing images. If there were much more images per individual, the success rates would be expected to become closer to each other; the results would be more stable.

Total Number of Images	**# of Training vs. # of Testing Images**	**Images used in Training**	**Images used in Testing**	**Success Rate (%)**
400	5-5	1,2,3,4,5	6,7,8,9,10	90,0
400	5-5	2,3,4,5,6	1,7,8,9,10	90,5
400	5-5	3,4,5,6,7	1,2,8,9,10	93,0
400	5-5	4,5,6,7,8	1,2,3,9,10	97,5
400	5-5	5,6,7,8,9	1,2,3,4,10	96,5
400	5-5	6,7,8,9,10	1,2,3,4,5	95,5
Average				93,8
400	9-1	1,2,3,4,5,6,7,8,9	10	95,0

Table 5.3: The success rates using different training and test images for each individual.

5.3.1.3. Using Different Number of Eigenvectors

In the third part of the experiments, the effect of using different number of eigenvectors while projecting the images (i.e. using Eigenfaces with different dimensions) is studied. In this stage, the tests are performed using both PCA and Subspace LDA (i.e. PCA+LDA).

From Figure 5.9, it can be observed that the performance of Subspace LDA is worse than that of PCA, when the eigenface space's dimension is small (35-60), and it is better than that of PCA, when the eigenface space's dimension is large (180-200). In fact, the decrease in the success rate, where the number of eigenvectors used to project the image is between 35 and 60, must be due to the fact that the eigenvectors can not linearly separate 40 classes. The number of eigenvectors should be sufficient to separate the classes used in training.

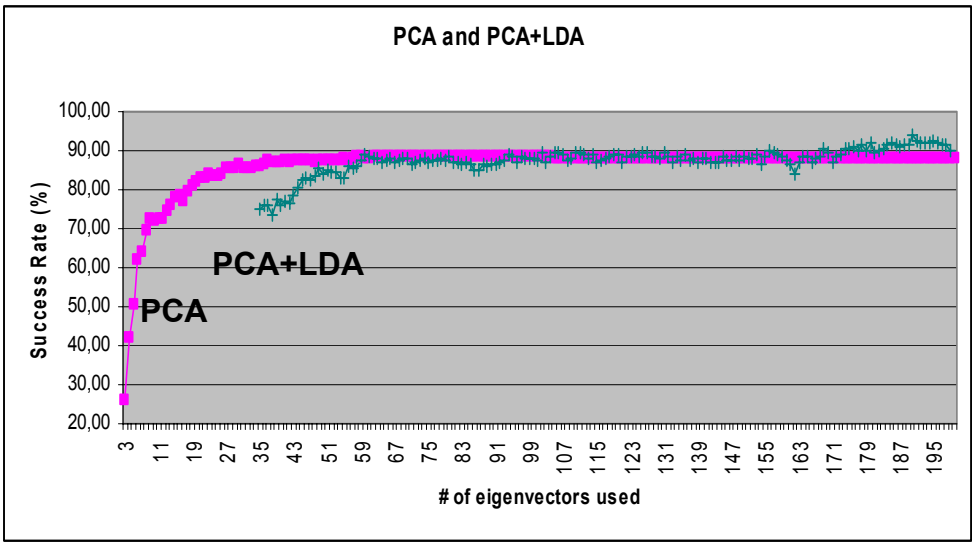


Figure 5.9: The success rates when using different number of eigenfaces.

5.3.1.4. Using Histogram Equalization and Rank-3

In this part of the experiments, the effect of using the preprocessing technique, histogram equalization and the rank-3 results of our experiments (i.e. the correct class being not in the first match, but in one of the first three matches) is studied. In this stage, the tests are performed using both PCA and Subspace LDA (i.e. PCA+LDA).

	Total Number of Images	**Number of Training vs. Number of Testing Images**	**Number of Eigenfaces used**	**Histeq**	**Rank**	**Success Rate (%)**
PCA	400	5-5	100	Yes	1	84,0
PCA	400	5-5	100	Yes	3	90,5
PCA	400	5-5	100	No	1	88,5
PCA	400	5-5	100	No	3	95,5
PCA	400	5-5	190	Yes	1	84,0
PCA	400	5-5	190	Yes	3	90,5
PCA	400	5-5	190	No	1	88,0
PCA	400	5-5	190	No	3	95,0
PCA + LDA	400	5-5	100	Yes	1	84,0
PCA + LDA	400	5-5	100	Yes	3	94,0
PCA + LDA	400	5-5	100	No	1	88,5
PCA + LDA	400	5-5	100	No	3	94,0
PCA + LDA	400	5-5	190	Yes	1	89,0
PCA + LDA	400	5-5	190	Yes	3	93,0
PCA + LDA	400	5-5	190	No	1	94,0
PCA + LDA	400	5-5	190	No	3	97,5

Table 5.4: The success rates when using histogram equalization and Rank-3.

From Table 5.4, it can be stated that the performance of PCA is not much influenced by using 190 eigenfaces instead of using 100 eigenfaces. As a result, using 100 dimensional eigenface space for PCA is sufficient. However, the performance of Subspace LDA increases

when 190 eigenfaces are used instead of 100 eigenfaces. This is an expected result, according to Figure 5.4. Applying histogram equalization decreases the performance in the ORL database. The decrease in performance when histogram equalization is applied is not clear, because in other databases (i.e. FERET and CNN-Turk Speakers Face Database) histogram equalization increases the performance.

Using rank-3 instead of rank-1 increases the performance between 2,5 % to 10 %. Thus, rank-3 may be useful for applications where possible top 3 matches are required instead of the top match. This may be used to reduce the number of possible individuals to match (e.g. the rank-3 results may be used as the previous step of a human operator based face recognition system).

5.3.1.5. Application of Pure LDA Method

In this part of the experiments, pure LDA method is implemented using different images sizes (given in Table 5.5).

	Image Dimensions	**Total Number of Images**	**Number of Training vs. Number of Testing Images**	**Number of Dimensions used**	**Success Rate (%)**
LDA	56 x 46	400	5-5	39	81,0
LDA	28 x 23	400	5-5	39	62,5
LDA	14 x 12	400	5-5	39	72,0

Table 5.5: The success rates when pure LDA is applied on different image sizes.

The important point in the LDA algorithm is that, the top *c-1* eigenvalues, *c* being the number of classes (40 in ORL), and their corresponding eigenvectors should be selected to obtain the classification

space. These eigenvalues become sometimes infinity and sometimes positive integers. Moreover, there is a disadvantage of the pure LDA algorithm: The application of the pure LDA algorithm on the ORL database of resized 56 x 46 images takes nearly 6-7 hours to complete. This complexity is unacceptable compared to other algorithms, which only take 3-5 minutes in average to complete the whole ORL database.

5.3.1.6. Application of the Bayesian PCA Method

In this part of the experiments, the Bayesian PCA method [24], [25], [26], [27], [28], [29], [30], [31] is implemented with different number of intrapersonal and extrapersonal eigenfaces. Moreover, different input image sizes as well as MAP and ML performances are also compared. Bayesian PCA method is implemented on face image differences, and there are two eigenface spaces as "extrapersonal eigenface space" and "intrapersonal eigenface space". Thus, this method is not compared against the PCA and Subspace LDA method; instead, their overall success rates are compared.

In Bayesian PCA, the computation of (N-M) as a power term in the equation of the likelihood (Equation 4.95) causes the equation to diverge (in Matlab, $\sim 10^{310}$ is assumed as infinity and $\sim 10^{-310}$ as zero). On the other hand, very small numbers return from the computation of $P(\Delta|\Omega_E)$, which may cause the overall probability value to go to zero. Thus, images are resized from 112 x 92 to 14 x 12 with a considerable recognition loss in these small sized images. Although the images are resized to such small dimensions, the results are comparable to that of Subspace LDA.

The results of the tests performed with ML are given in Table 5.6. The recognition rates of the MAP could not be calculated accurately, as

MAP includes computation of the $P(\Delta|\Omega_E)$ term which usually diverges in Matlab.

A disadvantage of this method is the computation time. It takes 2-3 hours to complete where other methods (except pure LDA which takes 6-7 hours to complete) take 3-5 minutes in average to complete the same number of test images, as each test image is differenced from each training image in the database to classify the image.

	Image Dimensions	Total Number of Images	Number of Training vs. Number of Test Images	Number of Intra-personal Eigenfaces	Number of Extra-personal Eigenfaces	Success Rate (%)
Bayesian PCA (ML)	14 x 12	400	4-6	70	70	90,8
Bayesian PCA (ML)	14 x 12	400	4-6	100	70	90,8
Bayesian PCA (ML)	14 x 12	400	4-6	50	70	92,9
Bayesian PCA (ML)	14 x 12	400	4-6	50	50	92,9
Bayesian PCA (ML)	14 x 12	400	4-6	40	50	93,3
Bayesian PCA (ML)	14 x 12	400	4-6	35	50	92,9
Bayesian PCA (ML)	14 x 12	400	4-6	30	50	93,3

Table 5.6: The success rates when different combinations are used in Bayesian PCA.

5.3.2. Tests Performed Using the FERET Face Database

Information about the standard FERET September 1996 gallery and probe sets is given in Table 5.1. As there is 1 image per individual in the FERET database standard gallery set, the tests are performed using

this set and different gallery and probe sets constructed using some other criteria. The studied test sets are as follows;

a) **The standard FERET dataset:** The files given in the FERET distribution CD of March 2001 are;

    1. *gallery.names*: contains 1196 images from 1196 individuals.
    2. *probe_fafb_expression.names*: contains 1195 images with alternative facial expressions.
    3. *probe_fafc_diffcamera_diffillum.names*: contains 194 images taken with different camera and under different illumination conditions.
    4. *probe_dup1_temporal_zero_to_34_months.names*: contains 722 images taken at anytime between the one minute and 34 months.
    5. *probe_dup2_at_least_18_months.names*: contains 234 images taken at anytime between 18 months and 34 months. This set is a subset of the dup1 set.

b) **The modified FERET dataset:** This dataset is constructed for this thesis using the individuals having more than 3 images in the database, especially to test the performance of Subspace LDA and Bayesian PCA, as they require at least 2 images per individual in the training phase. This dataset shows the performance of a system trained by images with different expressions, and tested for illumination and aging changes.

    1. *NewGalleryFaFb.names*: contains 1016 images from 508 individuals. First image of an individual comes from the

original FERET *gallery.names* file and the second image comes from the *probe_fafb_expression.names* file.

2. Same probe sets are used to test the systems, except for the *probe_fafb_expression.names* file as the images in this file are used in the training phase.

c) **The modified-2 FERET dataset:** This dataset is also constructed using the individuals having more than 3 images for the same purposes as explained in the previous set. This dataset shows the performance of a system trained by images taken at different times, and tested for expression and aging changes.

1. *NewGalleryFaDup1S.names*: contains 480 images from 240 individuals. The first image from an individual comes from the original FERET *gallery.names* file and the second image comes from the earliest images of the *probe_dup1_temporal_zero_to_34_months.names* file.

2. *probe_fafb_expressionMOD2.names*: contains 240 images constructed using alternative facial images which are corresponding mates of the gallery set in the *probe_fafb_expression.names* file.

3. Same probe sets are used to test the systems, except for the *probe_fafc_diffcamera_diffillum.names* file.

5.3.2.1. Using the Standard FERET Dataset

As there is 1 image per individual in the FERET database, only PCA could be tested with this dataset. The performance of the system after the preprocessing techniques cropping, histogram equalization and masking are applied on the image are given in Table 5.7. The images are resized to 50 x 50 before applying the preprocessing steps.

	Number of Training Images	Number of Testing Images	Number of Eigenfaces used	Applied Preprocessing	Success Rate (%)
PCA	1196 (gallery.names)	1195 (FaFb)	300	Crop + Histeq	74,4
PCA	1196 (gallery.names)	1195 (FaFb)	300	Crop + Histeq + Mask	74,6
PCA	1196 (gallery.names)	194 (FaFc)	300	Crop + Histeq	13,4
PCA	1196 (gallery.names)	194 (FaFc)	300	Crop + Histeq + Mask	9,8
PCA	1196 (gallery.names)	722 (Dup 1)	300	Crop + Histeq	29,7
PCA	1196 (gallery.names)	722 (Dup 1)	300	Crop + Histeq + Mask	31,0
PCA	1196 (gallery.names)	234 (Dup 2)	300	Crop + Histeq	10,3
PCA	1196 (gallery.names)	234 (Dup 2)	300	Crop + Histeq + Mask	12,0

Table 5.7: The success rates of the system after applying some of the preprocessing techniques.

Another set of tests are performed to see the effect of which preprocessing techniques applied and in which order they are applied. These tests are performed on the standard FaFb set of the FERET database, and the results are shown in Table 5.8. An important fact to point out here is that, the rotation technique applied in these tests is a very primitive one, which is developed in later sessions of the experiments.

The applied histogram equalization technique did not change throughout this study, but masking and rotation is developed in the next experiments which are performed with the modified FERET dataset and the modified-2 FERET dataset. The change in the masking and rotation techniques applied on the image, helps to much on putting the images in cannonical form which greatly enhances the recognition performance.

	Number of Training Images	Number of Testing Images	Number of Eigenfaces used	Applied Preprocessing	Success Rate (%)
PCA	1196 (gallery.names)	1195 (FaFb)	300	Rotate + Crop	66,8
PCA	1196 (gallery.names)	1195 (FaFb)	300	Crop	67,0
PCA	1196 (gallery.names)	1195 (FaFb)	300	Rotate + Crop + Histeq	74,6
PCA	1196 (gallery.names)	1195 (FaFb)	300	Crop + Histeq	74,4
PCA	1196 (gallery.names)	1195 (FaFb)	300	Rotate + Crop + Mask + Histeq	72,7
PCA	1196 (gallery.names)	1195 (FaFb)	300	Crop + Mask + Histeq	73,8
PCA	1196 (gallery.names)	1195 (FaFb)	300	Crop + Histeq + Mask	74,6
PCA	1196 (gallery.names)	1195 (FaFb)	200	Crop + Histeq	75,4
PCA	1196 (gallery.names)	1195 (FaFb)	200	Crop	62,6
PCA	1196 (gallery.names)	1195 (FaFb)	55	Crop	60,0

Table 5.8: The success rates of the system after applying the preprocessing techniques in different order and combinations.

5.3.2.2. Using the Modified FERET Dataset

Using the modified FERET dataset, constructed by the standard gallery and some of the images from the expression change probe set (i.e. FaFb) PCA, Subspace LDA and Bayesian LDA are tested in many combinations. All the three systems are trained using 2 images per individual. The test results are given in Tables 5.9, 5.10 and 5.11.

	Number of Training Images	Number of Testing Images	Number of Eigenfaces used	Applied Preprocessing	Image Dimensions	Success Rate (%)
PCA	1016 (FaFbMOD)	194 (FaFc)	300	Crop + Hist + Mask	70 x 70	18,0
PCA	1016 (FaFbMOD)	194 (FaFc)	300	Crop + Hist + Mask	50 x 50	13,0
PCA	1016 (FaFbMOD)	194 (FaFc)	300	Crop + Hist + Mask	20 x 20	17,5
PCA	1016 (FaFbMOD)	194 (FaFc)	300	Rot + Crop + Hist + Mask	70 x 70	15,5
PCA	1016 (FaFbMOD)	194 (FaFc)	300	Rot + Crop + Hist + Mask	50 x 50	16,5
PCA	1016 (FaFbMOD)	722 (Dup 1)	300	Crop + Hist + Mask	70 x 70	38,9
PCA	1016 (FaFbMOD)	722 (Dup 1)	300	Crop + Hist + Mask	50 x 50	38,4
PCA	1016 (FaFbMOD)	722 (Dup 1)	300	Crop + Hist + Mask	20 x 20	37,1
PCA	1016 (FaFbMOD)	722 (Dup 1)	300	Rot + Crop + Hist + Mask	70 x 70	43,4
PCA	1016 (FaFbMOD)	722 (Dup 1)	300	Rot + Crop + Hist + Mask	50 x 50	42,4
PCA	1016 (FaFbMOD)	234 (Dup 2)	300	Crop + Hist + Mask	70 x 70	17,5
PCA	1016 (FaFbMOD)	234 (Dup 2)	300	Crop + Hist + Mask	50 x 50	16,7
PCA	1016 (FaFbMOD)	234 (Dup 2)	300	Crop + Hist + Mask	20 x 20	18,0
PCA	1016 (FaFbMOD)	234 (Dup 2)	300	Rot + Crop + Histeq + Mask	70 x 70	18,8
PCA	1016 (FaFbMOD)	234 (Dup 2)	300	Rot + Crop + Hist + Mask	50 x 50	18,8

Table 5.9: The performance of PCA on modified FERET datataset.

	Number of Training Images	Number of Testing Images	Number of Eigenfaces used	Applied Preprocessing	Image Dimensions	Success Rate (%)
PCA+ LDA	1016 (FaFbMOD)	194 (FaFc)	300	Crop + Hist + Mask	70 x 70	27,3
PCA+ LDA	1016 (FaFbMOD)	194 (FaFc)	300	Crop + Hist + Mask	50 x 50	28,9
PCA+ LDA	1016 (FaFbMOD)	194 (FaFc)	300	Crop + Hist + Mask	20 x 20	18,6
PCA+ LDA	1016 (FaFbMOD)	194 (FaFc)	300	Rot + Crop + Hist + Mask	70 x 70	24,2
PCA+ LDA	1016 (FaFbMOD)	194 (FaFc)	300	Rot + Crop + Hist + Mask	50 x 50	28,9
PCA+ LDA	1016 (FaFbMOD)	722 (Dup 1)	300	Crop + Hist + Mask	70 x 70	39,2
PCA+ LDA	1016 (FaFbMOD)	722 (Dup 1)	300	Crop + Hist + Mask	50 x 50	40,3
PCA+ LDA	1016 (FaFbMOD)	722 (Dup 1)	300	Crop + Hist + Mask	20 x 20	36,0
PCA+ LDA	1016 (FaFbMOD)	722 (Dup 1)	300	Rot + Crop + Hist + Mask	70 x 70	47,1
PCA+ LDA	1016 (FaFbMOD)	722 (Dup 1)	300	Rot + Crop + Hist + Mask	50 x 50	49,0
PCA+ LDA	1016 (FaFbMOD)	234 (Dup 2)	300	Crop + Hist + Mask	70 x 70	29,9
PCA+ LDA	1016 (FaFbMOD)	234 (Dup 2)	300	Crop + Hist + Mask	50 x 50	32,9
PCA+ LDA	1016 (FaFbMOD)	234 (Dup 2)	300	Crop + Hist + Mask	20 x 20	30,3
PCA+ LDA	1016 (FaFbMOD)	234 (Dup 2)	300	Rot + Crop + Hist + Mask	70 x 70	41,9
PCA+ LDA	1016 (FaFbMOD)	234 (Dup 2)	300	Rot + Crop + Hist + Mask	50 x 50	41,5

Table 5.10: The performance of Subspace LDA (PCA + LDA) on modified FERET dataset.

	Number of Training Images	Number of Testing Images	Number of Eigenfaces used	Applied Preprocessing	Image Dimensions	Success Rate (%)
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	400 (150-150)	Crop + Hist + Mask	30 x 30	28,9
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	400 (200-200)	Crop + Hist + Mask	30 x 30	23,2
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (120-90)	Crop + Hist + Mask	30 x 30	9,8
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (300-300)	Crop + Hist + Mask	20 x 20	26,8
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (150-150)	Crop + Hist + Mask	20 x 20	28,4
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (100-100)	Crop + Hist + Mask	20 x 20	27,3
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (120-90)	Crop + Hist + Mask	22 x 22	26,3
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (90-120)	Crop + Hist + Mask	22 x 22	25,3
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	400 (150-150)	Rot + Crop + Hist + Mask	22 x 22	24,2
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (150-150)	Rot + Crop + Hist + Mask	50 x 50	24,2
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (150-150)	Rot + Crop + Hist + Mask	30 x 30	24,2
Bayesian PCA (MAP)	1016 (FaFbMOD)	194 (FaFc)	300 (120-90)	Rot + Crop + Hist + Mask	22 x 22	28,9
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	400 (150-150)	Crop + Hist + Mask	30 x 30	41,7
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (300-300)	Crop + Hist + Mask	20 x 20	33,4
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (150-150)	Crop + Hist + Mask	20 x 20	37,5
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (120-90)	Crop + Hist + Mask	22 x 22	34,2
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (90-120)	Crop + Hist + Mask	22 x 22	40,4
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	400 (150-150)	Rot + Crop + Hist + Mask	22 x 22	35,2
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (150-150)	Rot + Crop + Hist + Mask	50 x 50	50,0
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (120-90)	Rot + Crop + Hist + Mask	50 x 50	50,4

Table 5.11: The performance of Bayesian PCA (MAP) on modified FERET dataset.

	Number of Training Images	Number of Testing Images	Number of Eigenfaces used*	Applied Preprocessing	Image Dimension	Success Rate (%)
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (90-120)	Rot + Crop + Hist + Mask	50 x 50	50,0
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (150-150)	Rot + Crop + Hist + Mask	30 x 30	46,4
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (120-90)	Rot + Crop + Hist + Mask	22 x 22	42,8
Bayesian PCA (MAP)	1016 (FaFbMOD)	722 (Dup 1)	300 (90-120)	Rot + Crop + Hist + Mask	22 x 22	48,2
Bayesian PCA (MAP)	1016 (FaFbMOD)	234 (Dup 2)	300 (300-300)	Crop + Hist + Mask	20 x 20	26,9
Bayesian PCA (MAP)	1016 (FaFbMOD)	234 (Dup 2)	300 (120-90)	Crop + Hist + Mask	22 x 22	29,1
Bayesian PCA (MAP)	1016 (FaFbMOD)	234 (Dup 2)	300 (90-120)	Crop + Hist + Mask	22 x 22	29,5
Bayesian PCA (MAP)	1016 (FaFbMOD)	234 (Dup 2)	400 (150-150)	Rot + Crop + Hist + Mask	22 x 22	32,1
Bayesian PCA (MAP)	1016 (FaFbMOD)	234 (Dup 2)	300 (150-150)	Rot + Crop + Hist + Mask	50 x 50	34,2
Bayesian PCA (MAP)	1016 (FaFbMOD)	234 (Dup 2)	300 (150-150)	Rot + Crop + Hist + Mask	30 x 30	35,0
Bayesian PCA (MAP)	1016 (FaFbMOD)	234 (Dup 2)	300 (120-90)	Rot + Crop + Hist + Mask	22 x 22	33,8

(* In the "Number of eigenfaces used" field first number means the total number of eigenfaces used, the first number in the parenthesis gives the number of the intrapersonal eigenfaces used, the last number gives the extrapersonal eigenfaces used in the recognition process.)

Table 5.11 (cont.): The performance of Bayesian PCA (MAP) on modified FERET dataset.

In Bayesian PCA (MAP), many tests have to be conducted on the system in order the system to converge. The system converges in the MAP case, whereas diverges in the ML case.

By examining the results of these tests, PCA does its best performance on the FaFc (illumination change) probe set with 18,0 %,

which is slightly small compared to the performance of Subspace LDA (28,9 % on the same probe set) and Bayesian PCA (MAP) does its best at 28,9 %, exactly same as the Subspace LDA. By considering these results, it can be argued that Bayesian PCA and Subspace LDA perform almost the same under different illumination conditions, whereas PCA performes slightly worse for this case. However, the best performance rate which is 28,9 % is a low recognition rate, which states that all the systems under test perform inferior under different illumination conditions.

Moreover, by looking at the results of the Dup 1 test on the modified FERET set, PCA performed 43,4 %, again which is slightly less than the performance of Subspace LDA (49,0 %). Bayesian PCA (MAP) performed 50,4 % which is quite near to the performance of Subspace LDA. When these results are evaluated, again Subspace LDA and Bayesian PCA perform slightly better that PCA under small time change, but again the performance is not fair for all the systems under test.

According to the results of the Dup 2 test on the modified FERET set, PCA performed 18,8 % which is in this case very low compared to the performance of Subspace LDA (41,9 %), and the best performance Bayesian PCA (MAP) obtained for this set is 35,0 %. According to the results of this test, Subspace LDA performs best under aging, Bayesian PCA does slightly worse compared to Subspace LDA; PCA performs the worst.

Moreover, when all the cases are focused well, it can be easily argued that using higher dimension images does not help the system too much, perhaps it causes overfitting.

5.3.2.3. Using the Modified-2 FERET Dataset

Based on the experience from the previous experiments, modified-2 FERET dataset is tested under the best conditions obtained in the standard FERET dataset and the modified FERET dataset.

Table 5.12 and Table 5.13 contain the test results performed on the modified-2 FERET dataset with different image sizes and different number of intrapersonal and extrapersonal eigenvectors used.

	Number of Training Images	**Number of Testing Images**	**Number of Eigenfaces used**	**Applied Preprocessing**	**Image Dimensions**	**Success Rate (%)**
PCA	480 (FaDup1)	240 (FaFb MOD)	300	Rot + Crop + Hist + Mask	70 x 70	83,3
PCA	480 (FaDup1)	240 (FaFb MOD)	300	Rot + Crop + Hist + Mask	50 x 50	83,3
PCA	480 (FaDup1)	479 (Dup1 MOD2)	300	Rot + Crop + Hist + Mask	70 x 70	57,8
PCA	480 (FaDup1)	479 (Dup1 MOD2)	300	Rot + Crop + Hist + Mask	50 x 50	57,2
PCA	480 (FaDup1)	234 (Dup 2)	300	Rot + Crop + Hist + Mask	70 x 70	49,1
PCA	480 (FaDup1)	234 (Dup 2)	300	Rot + Crop + Hist + Mask	50 x 50	48,3
PCA+ LDA	480 (FaDup1)	240 (FaFb MOD)	230	Rot + Crop + Hist + Mask	70 x 70	73,8
PCA+ LDA	480 (FaDup1)	240 (FaFb MOD)	230	Rot + Crop + Hist + Mask	50 x 50	80,8
PCA+ LDA	480 (FaDup1)	479 (Dup1 MOD2)	230	Rot + Crop + Hist + Mask	70 x 70	72,7
PCA+ LDA	480 (FaDup1)	479 (Dup1 MOD2)	230	Rot + Crop + Hist + Mask	50 x 50	77,5
PCA+ LDA	480 (FaDup1)	234 (Dup 2)	230	Rot + Crop + Hist + Mask	70 x 70	69,7
PCA+ LDA	480 (FaDup1)	234 (Dup 2)	230	Rot + Crop + Hist + Mask	50 x 50	73,1

Table 5.12: The performance of PCA, Subspace LDA on modified-2 FERET dataset with different image sizes and different number of eigenfaces used.

	Number of Training Images	Number of Testing Images	Number of Eigenfaces used	Applied Preprocessing	Image Dimensions	Success Rate (%)
MAP	480 (FaDup1)	240 (FaFb MOD)	230 (150-150)	Rot + Crop + Hist + Mask	30 x 30	96,3
MAP	480 (FaDup1)	240 (FaFb MOD)	230 (120-90)	Rot + Crop + Hist + Mask	30 x 30	97,1
MAP	480 (FaDup1)	240 (FaFb MOD)	230 (90-120)	Rot + Crop + Hist + Mask	30 x 30	96,7
MAP	480 (FaDup1)	240 (FaFb MOD)	230 (120-90)	Rot + Crop + Hist + Mask	22 x 22	97,1
MAP	480 (FaDup1)	240 (FaFb MOD)	230 (90-120)	Rot + Crop + Hist + Mask	22 x 22	95,0
MAP	480 (FaDup1)	479 (Dup1 MOD2)	230 (150-150)	Rot + Crop + Hist + Mask	30 x 30	87,1
MAP	480 (FaDup1)	479 (Dup1 MOD2)	230 (120-90)	Rot + Crop + Hist + Mask	30 x 30	87,7
MAP	480 (FaDup1)	479 (Dup1 MOD2)	230 (90-120)	Rot + Crop + Hist + Mask	30 x 30	87,1
MAP	480 (FaDup1)	479 (Dup1 MOD2)	230 (120-90)	Rot + Crop + Hist + Mask	22 x 22	85,8
MAP	480 (FaDup1)	479 (Dup1 MOD2)	230 (90-120)	Rot + Crop + Hist + Mask	22 x 22	85,0
MAP	480 (FaDup1)	234 (Dup 2)	230 (150-150)	Rot + Crop + Hist + Mask	30 x 30	82,1
MAP	480 (FaDup1)	234 (Dup 2)	230 (120-90)	Rot + Crop + Hist + Mask	30 x 30	83,3
MAP	480 (FaDup1)	234 (Dup 2)	230 (90-120)	Rot + Crop + Hist + Mask	30 x 30	81,6
MAP	480 (FaDup1)	234 (Dup 2)	230 (120-90)	Rot + Crop + Hist + Mask	22 x 22	79,9
MAP	480 (FaDup1)	234 (Dup 2)	230 (90-120)	Rot + Crop + Hist + Mask	22 x 22	79,9

Table 5.13: The performance Bayesian PCA (MAP) on modified-2 FERET dataset with different image sizes and different number of eigenfaces used.

The results of the modified-2 FERET dataset show that, the best performance is obtained in the Bayesian PCA (fafbMOD 97,1 %, dup1MOD2 87,1 %, dup2 83,3 %). Subspace LDA performs better than PCA in dup1MOD2 (Subspace LDA 72,7 %, PCA 57,8 %) and dup2

(Subspace LDA 73,1 %, PCA 49,1 %), whereas PCA performs slightly better than Subspace LDA in fafbMOD (Subspace LDA 80,8 %, PCA 83,3 %).

The best success rates are obtained for all the systems under test using modified-2 FERET dataset. Thus, training these systems with the images of an individual taken at different times results with good classification performance.

5.3.3. Tests Performed Using the CNN-TURK Speakers Face Database

Using the CNN-TURK Speakers Face Database, 2 tests are conducted on PCA and Subspace LDA. The results are given in Table 5.14. The 5-13 training vs. test image set includes images captured on the same day, whereas 5-5 training vs. test image set includes test images captured 7-10 days after the training images are captured.

	Total Number of Images	Number of Training vs. Number of Testing Images	Number of Eigenfaces used	Histeq	Success Rate (%)
PCA	198 (CNN)	5-13	55	No	99,3
PCA	198 (CNN)	5-13	55	Yes	99,3
PCA	110 (CNN)	5-5	55	No	58,2
PCA	110 (CNN)	5-5	55	Yes	87,3
PCA + LDA	198 (CNN)	5-13	55	No	98,6
PCA + LDA	198 (CNN)	5-13	55	Yes	99,3
PCA + LDA	110 (CNN)	5-5	55	No	74,5
PCA + LDA	110 (CNN)	5-5	55	Yes	96,4

Table 5.14: The success rates when using histogram equalization on CNN-TURK Speakers face database.

According to these results, the recognition performance is quite good for the images captured on the same day for both PCA and Subspace LDA. However, performance decreases comparably (58,2 % for PCA and 74,5 % for Subspace LDA) for the images taken 7-10 days later, where especially there is illumination change and change in hair style or pose, etc. This performance decrease can be recovered (87,3 % for PCA and 96,4 % for Subspace LDA) using histogram equalization for both of the algorithms. Again the tests performed on this database shows that PCA is more susceptible to such changes compared to Subspace LDA. Although the changes are recovered using histogram equalization, PCA could not reach its initial performance whereas Subspace LDA mostly could get well.

Moreover, the tests performed with this database shows that both PCA and Sabspace LDA can handle the noise due to interleaving and other noises in images captured from video.

CHAPTER 6

CONCLUSIONS

This thesis study mainly focuses on the holistic face recognition approaches. Throughout the study, the performance of the statistical approaches such as Principal Component Analysis (PCA), Subspace Linear Discriminant Analysis (Subspace LDA) and Bayesian PCA are investigated. PCA and LDA are used to obtain the features from the given face images, and Bayesian Classifier and Nearest Mean Classifier are used to classify the face images using these features.

The investigated techniques are not novel techniques; they are previously proposed and studied in the literature. PCA was proposed in 1991, and based on PCA, Bayesian PCA and Subspace LDA were later proposed. They were also tested under different sets of face databases. The main purpose is studying the performance of Subspace LDA and Bayesian PCA methods in depth, which are two competeting methods reported to be among the best performing methods in the literature.

Hence, it is better to get a first hand experience on these approaches and understand the power and weakness of them. The methods are applied onto 3 databases, ORL (Olivetti Research Laboratory) database, FERET (Face Recognition Technology) database and a newly constructed database throughout this study, the CNN-TURK Speakers Face Database.

ORL database is the first face database that the performance of the methods was tested. PCA is applied on to the ORL database first to decide the number of training and test images to use in order to obtain better recognition rates. Among different combinations of training and test images, the best recognition is obtained under 5 training and 5 testing images case. Hence, in the rest of the study, ORL is tested with 5 images in training set and 5 images in testing set.

In the next phase of experiments, the effect of changing the training and test sets for the same individual is studied using Subspace LDA. The recognition rate is 93,8 ± 3.8 % for different sets of training and test images. It is observed that changing the test and probe sets slightly affects the performance of the method under test. Hence, there is no need to study the performance change with respect to the change in the training and probe sets, the same test sets are used in ORL throughout the study.

The effect of using different number of eigenvectors (i.e. different eigenface space size) is also tested in PCA and Subspace LDA. It is observed that the performance of Subspace LDA is inferior compared to PCA when small number of eigenfaces is used; however, its performance is better compared to PCA when most of the eigenfaces are used. Although, the performance of PCA always increases while increasing the number of eigenvectors, the performance of Subspace LDA starts to decrease after 190 eigenfaces (for a total 200 eigenfaces). It can be argued that overfitting causes this problem; in other words, the system starts to memorize after using 190 eigenfaces.

Moreover, as a further study, pure LDA is also performed on ORL database with different image sizes. Although, its performance is 81 %, which is relatively acceptable, the execution time is long due to its

computational complexity. Hence, pure LDA is not applied on any other database.

Bayesian PCA method is implemented with different number of intrapersonal and extrapersonal eigenfaces, different image sizes used as input data. Moreover, two variances of Bayesian PCA; Maximum A Posteriori (MAP) and Maximum Likelihood (ML) performance are compared on the ORL database. The recognition performance of ML is better than MAP performance in the ORL database. The reason for this is due to the computation of $P(\Delta|\Omega_E)$ term in MAP, which usually diverges, causing bad recognition performance. ML performs its best with 93,3 % when the images are resized from 112 x 92 to 14 x 12, 4 training and 6 test images are used, and 30 (or 40) intrapersonal eigenfaces and 50 extrapersonal eigenfaces are used.

In Bayesian PCA, the dimension of the resized image is critical: If it is too large, the system may diverge, whereas if it is too small, the images cannot be discriminated from each other. Moreover, since image differences are used in training (in this method) and each test image is subtracted from each training image to observe if it belongs to that individual, it has higher computational complexity.

The CNN-TURK Speakers Face Database is the next database where only the performance of PCA and Subspace LDA are tested. The database tests are designed to observe the effect of pose change, illumination change, 7-10 days aging, and the noise coming from the images captured from video, especially due to interlacing. Both methods perform quite well for pose changes. The performance of PCA is affected by illumination change, and 7-10 days of aging, but this performance decrease can be significantly handled by simple histogram equalization. On the other hand, the performance of Subspace LDA is not much

susceptible to these changes. The system performs quite well (96,4 %) using histogram equalization.

Moreover, both PCA and Subspace LDA performed quite well in CNN-TURK Speakers face database, where there is an extra noise coming from the images due to interlacing; thus, it can be argued that both methods are successful in handling the noises coming out of interlacing.

FERET database is the final database used during simulations. FERET images are used in three different ways. As there is one image per individual in the standard FERET training set, two other modifications of the standard FERET subsets are produced.

The standard FERET dataset is tested on PCA. The effects of the preprocessing techniques cropping, histogram equalization, and masking are observed. The preprocessing techniques are important since they put the images in a canonical form, which greatly enhances the face recognition performance. From the results it can be seen that the following preprocessing combinations performed superior: (Rotation + Cropping + Histogram Equalization), (Cropping + Histogram Equalization + Masking), and (Cropping + Histogram Equalization).

Next, the modified FERET dataset is constructed by combining the gallery images and alternative expression images (fafb) and tested under different illumination conditions (fafc), aging from 1 minute to 34 months (dup1), and aging from 18 months to 34 months (dup2).

The results of the fafc set show that best performance of PCA (18,0 %) is slightly lower compared to the best performances of Subspace LDA and Bayesian PCA, which have the same performance (28,9 %). However, it can be argued that all the algorithms under test

perform quite low when illumination direction or strength changes from image to image.

According to the results obtained in the dup1 set, again the performances of Subspace LDA (49,0 %) and Bayesian PCA (50,4 %) are slightly better to that of PCA (43,4 %). In fact, again the performances of all the three systems under test in the dup1 set are not satisfactory.

Next, according to the results of the dup2 set, PCA performed inferior (18,8 %) with respect to Subspace LDA (41,9 %) and Bayesian PCA (35,0 %). Subspace LDA gives the best results under aging, the performance of Bayesian PCA is slightly worse than Subspace LDA, and PCA has the worst performance.

Moreover, when all the test results performed on the modified FERET dataset are compared, it can be argued that higher dimension does not mean better classification where overfitting occurs and computational complexity increases.

Finally, the modified-2 FERET dataset is constructed by taking the first image of each individual in the dup1 set and the corresponding image of the same individual in the standard FERET gallery set. This system is tested under facial expression change (of the images contained in the gallery, fafbMOD), the rest of the dup1 probe set (dup1MOD2), and the whole dup2 probe set.

According to the results of the modified-2 FERET dataset, the best performance is obtained by Bayesian PCA (fafbMOD 97,1 %, dup1MOD2 87,1 %, dup2 83,3 %). Subspace LDA performs better than PCA in dup1MOD2 (Subspace LDA 72,7 %, PCA 57,8 %) and dup2 (Subspace LDA 73,1 %, PCA 49,1 %), whereas PCA performs better

than Subspace LDA in fafbMOD (Subspace LDA 80,8 %, PCA 83,3 %). This relative decrease in the performance of Subspace LDA (with respect to PCA and Bayesian PCA) must be due to the fact that classes cannot be linearly separated because of the large within class scatter of each class. Consequently, the image classes can not be separated from each other.

Moreover, comparing the results of the tests applied on modified FERET dataset with modified-2 FERET dataset, it can be suggested that using the images of an individual taken at different times (in this case from 1 minute to 34 months) for training gives better performance results compared to using the images of an individual taken a few seconds apart. This is due to the variance added to the training set by aging effects. Thus, the systems are able to compensate for the varieties of each face class.

There is a tradeoff at this point: Should the variance be small for the Subspace LDA to easily discriminate each face class or should it be big for the system to be able to compensate for the varieties of each individual (i.e. should the training images be selected in order to obtain the biggest variance or should they be selected to obtain relatively small variance, if possible)? It can be argued that it is better to use more training images with Subspace LDA in order to obtain large eigenface space dimensions for the Subspace LDA discrimination when there is huge variation within each class. In the absence of more training images it would be better to use Bayesian PCA which shows the best performance when there is much variance.

As a result, it can be argued that both Subspace LDA and Bayesian PCA perform quite successfully (difficult to argue that one of them is better than the other) and the performance for PCA is slightly worse than that of Subspace LDA and Bayesian PCA. Using an extra

classifier after PCA, such as LDA or modeling the distance between PCA outputs in a Bayesian formulation (i.e. Bayesian PCA) brings this enhanced performance.

REFERENCES

[1] P. S. Huang, C. J. Harris, and M. S. Nixon, "Human Gait Recognition in Canonical Space using Temporal Templates", IEE Proc.-Vis. Image Signal Process, Vol. 146, No. 2, April 1999.

[2] J. Huang, "Detection Strategies For Face Recognition Using Learning and Evolution", PhD. Thesis, George Mason University, May 1998.

[3] S.A. Rizvi, P.J. Phillips, and H. Moon, "A verification protocol and statistical performance analysis for face recognition algorithms", pp. 833-838, IEEE Proc. Conf. Computer Vision and Pattern Recognition (CVPR), Santa Barbara, June 1998.

[4] R. Chellappa, C. L. Wilson and S. Sirohey, "Human and Machine Recognition of Faces: A Survey", Proceedings of the IEEE, Vol. 83, No. 5, May 1995.

[5] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face Recognition: A literature Survey", Technical Report, Univ. of Maryland, 2000.

[6] J. J. Weng and D. L. Swets, "Face Recognition", in A. K. Jain, R. Bolle, and S. Pankanti (Editors), Biometrics: Personal Identification in Networked Society, Kluwer Academic Press, 1999.

[7] V. Bruce, "Identification of Human Faces", pp. 615-619, Image Processing and Its Applications, Conference Publication No. 465, IEEE, 1999.

[8] P. Temdee, D. Khawparisuth, and K. Chamnongthai, "Face Recognition by Using Fractal Encoding and Backpropagation Neural Network", pp. 159-161, 5th International Symposium on Signal Processing and its Applications, ISSPA '99, Australia, August 1999.

[9] R. Brunelli, and T. Poggio, "Face Recognition: Features versus Templates", pp. 1042-1052, IEEE Transactions, PAMI, 15(10). 1993.

[10] M. A. Turk, and A. P. Pentland, "Face Recognition using Eigenfaces", pp. 586-591, IEEE, 1991.

[11] L. Sirovich, and M. Kirby, "Low-dimensional Procedure for the Characterization of Human Faces", pp. 519-524, Journal of the Optical Society of America, Vol. 4, No. 3, March 1987.

[12] M. Kirby, and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces", pp. 103-108, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No. 1, January 1990.

[13] B. Kepenekci, "Face Recognition Using Gabor Wavelet Transform", MSc. Thesis, METU, September 2001.

[14] B. Moghaddam, and A. Pentland, "An Automatic System for Model-Based Coding of Faces", pp. 362-370, IEEE, 1995.

[15] S. J. Lee, S. B. Yung, J. W. Kwon, and S. H. Hong, "Face Detection and Recognition Using PCA", pp. 84-87, IEEE TENCON, 1999.

[16] S. Z. Lee, and J. Lu, "Generalizing Capacity of Face Database for Face Recognition", pp. 402-406, IEEE, 1998.

[17] J. L. Crowley, and K. Schwerdt, "Robust Tracking and Compression for Video Communication", pp. 2-9, IEEE, 1999.

[18] S. Cagnoni, A. Poggi, "A Modified Modular Eigenspace Approach to Face Recognition", pp. 490-495, IEEE, 1999.

[19] K. Etemad, and R. Chellappa, "Face Recognition Using Discriminant Eigenvectors", pp. 2148-2151, IEEE, 1996.

[20] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Spesific Linear Projection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, July 1997.

[21] W. Zhao, A. Krishnaswamy, R. Chellappa, D. L. Swets, and J. Weng, "Discriminant Analysis of Principal Components for Face Recognition", pp. 336-341, International Conference on Automatic Face and Gesture Recognition, 1998.

[22] W. Zhao, R. Chellappa, and N. Nandhakumar, "Emprical Performance Analysis of Linear Discriminant Classifiers", pp. 164-169, IEEE, 1998.

[23] W. Zhao, "Subspace Methods in Object/Face Recognition", pp. 3260-3264, IEEE, 1999.

[24] B. Moghaddam, "Principal Manifolds and Bayesian Subspaces for Visual Recognition", pp. 1131-1135, IEEE, 1999.

[25] B. Moghaddam, and A. Pentland, "Bayesian Image Retrieval in Biometric Databases", pp. 610-615, IEEE, 1999.

[26] B. Moghaddam, T. Jebara, and A. Pentland, "Efficient MAP/ML Similarity Matching for Visual Recognition", pp. 876-881, IEEE, 1998.

[27] B. Moghaddam, and A. Pentland, "Beyond Euclidean Eigenspaces: Bayesian Matching for Visual Recognition", Face Recognition: From Theories to Applications, pp. 921-930, October 1998.

[28] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian Face Recognition", pp. 1771-1782, Pattern Recognition, Vol. 33, No. 11, November 2000.

[29] B. Moghaddam, and A. Pentland, "Probabilistic Visual Learning for Object Detection", The 5th International Conference on Computer Vision, Cambridge, MA, June 1995.

[30] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian Modelling of Facial Similarity", Advances in Neural Information Processing Systems 11, MIT Press, 1999.

[31] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond Eigenfaces: Probabilistic Matching for Face Recognition", The 3rd International Conference on Automatic Face & Gesture Recognition, Nara, Japan, April 1998.

[32] B. Moghaddam, and A. Pentland, "Probabilistic Visual Learning for Object Representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, July 1997.

[33] B. Moghaddam, "Principal Manifolds and Probabilistic Subspaces for Visual Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 6, June 2002.

[34] C. Liu, and H. Wechsler, "A Unified Bayesian Framework for Face Recognition", pp. 151-155, IEEE, 1998.

[35] C. Liu, and H. Wechsler, "Probabilistic Reasoning Models for Face Recognition", pp. 827-832, IEEE, 1998.

[36] K. C. Chung, S. C. Kee, and S. R. Kim, "Face Recognition using Principal Component Analysis of Gabor Filter Responses", p. 53-57, IEEE, 1999.

[37] C. Podilchuk, and X. Zhang, "Face Recognition Using DCT-Based Feature Vectors", pp. 2144-2147, IEEE, 1996.

[38] S. Eickeler, S. Müller, and G. Rigoll, "High Quality Face Recognition in JPEG Compressed Images", pp. 672-676, IEEE, 1999.

[39] A. V. Nefian, and M. H. Hayes, "Hidden Markow Models for Face Recognition", pp. 2721-2724, IEEE, 1998.

[40] P. J. Phillips, "Support Vector Machines Applied to Face Recognition", pp. 803-809, Advances in Neural Information Processing Systems 11, MIT Press, 1999.

[41] H. Spies, and I. Ricketts, "Face Recognition in Fourier Space", Vision Interface '2000: 38-44, Montreal, 2000.

[42] C. S. Bobis, R. C. Gonzalez, J. A. Cancelas, I. Alvarez, and J. M. Enguita, "Face Recognition Using Binary Thresholding for Features Extraction", pp. 1077-1080, IEEE, 1999.

[43] A. X. Guan, and H. H. Szu, "A Local Face Statistics Recognition Methodology beyond ICA and/or PCA", pp. 1016-1027, IEEE, 1999.

[44] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", John Wiley & Sons, 2nd Edition, 2001.

[45] A. Martinez, "Face Image Retrieval Using HMMs", pp. 35-39, IEEE, 1999.

[46] P. Temdee, D. Khawparisuth, and K. Chamnongthai, "Face Recognition by Using Fractal Encoding and Backpropagation Neural Network", pp. 159-161, Fifth International Symposium on Signal Processing and its Applications, ISSPA '99, Brisbane, Australia, August 1999.

[47] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. Von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture", pp. 300-310, IEEE Transactions on Computers, Vol. 42, March 1993.

[48] L. Wiskott, J. M. Fellous, N. Krüger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching", pp. 129-132, IEEE, 1997.

[49] M. J. Er, S. Wu, and J. Lu, "Face Recognition Using Radial Basis Function (RBF) Neural Networks", Proceedings of the 38th Conference on Decision & Control, Phoenix, Arizona USA, pp. 2162-2167, IEEE, 1999.

[50] C. E. Thomaz, R. Q. Feitosa, and A. Veiga, "Design of Radial Basis Function Network as Classifier in Face Recognition Using Eigenfaces", pp. 118-123, IEEE, 1998.

[51] Y. Yoshitomi, T. Miyaura, S. Tomito, and S. Kimura, "Face Identification Using Thermal Image Processing", pp. 374-379, IEEE International Workshop on Robot and Human Communication, IEEE, 1997.

[52] Z. Liposcak, and S. Loncaric, "Face Recognition from Profiles Using Morphological Operations", pp. 47-52, IEEE, 1999.

[53] R. J. Schalkoff, "Pattern Recognition", John Wiley & Sons, 1992.

[54] E. Taslidere, "Face Detection Using a Mixture of Subspaces", MSc. Thesis, METU, September 2002.

[55] Olivetti & Oracle Research Laboratory, The Olivetti & Oracle Research Laboratory Face Database of Faces, http://www.camorl.co.uk/facedatabase.html

[56] P. J. Phillips, H. Moon, S.A. Rizvi, and P. J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms", pp. 1090-1104, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10, October 2000.

[57] P. J. Phillips, P. J. Rauss, and S. Z. Der, "FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results", Army Research Lab Technical Report 995, October 1996.

[58] P. J. Phillips, and H. Moon, "Comparison of Projection-Based Face Recognition Algorithms", p. 4057-4062, IEEE, 1997.