Feature regularization and extraction in eigenspace for face recognition

Bappaditya Mandal

2008

Bappaditya, M. (2008). Feature regularization and extraction in eigenspace for face recognition. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/13278>

<https://doi.org/10.32657/10356/13278>

Feature Regularization and Extraction in Eigenspace for Face Recognition

Bappaditya Mandal

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University

in fulfillment of the requirements for the degree of

Doctor of Philosophy

2008

Statement of Originality

I hereby certify that the content of this thesis is the result of work done by me and has not been submitted for a higher degree to any other University or Institution.

.....

Date

.....

Bappaditya Mandal

Acknowledgments

I would like to express my deep and sincere gratitude to my supervisor, Prof. Xudong Jiang. His enthusiastic engagement in my research and his never ending stream of ideas has been absolutely essential for the results presented here. I am very grateful that he has spent so much time with me discussing different problems ranging from philosophical issues down to minute technical details. I am grateful to my co-supervisor Prof. Alex Kot for his valuable comments and discussions during the development of algorithms. Many thanks to Prof. Xudong Jiang and Prof. Alex Kot, whom I have always admired not only for their remarkable scientific achievements, but also their kindness for helping young researchers like me.

I would like to thank Prof. Yap Kim Hui, Prof. Farook Sattar, Prof. Anamitra Makur, Prof. Chan Kap Luk and Prof. Sirajudeen Razul for their active help and guidance in understanding many fundamental concepts. I would also like to thank all my friends in workstation resource laboratory, which includes Niladri, Manhua Liu, Yap Pew Thian, Gao Yupeng, Yang Huijuan, Cheng Jun and Cao Hong as well as Stewart and Dorothy, who have been very supportive over these years. Outside my laboratory, I would like to thank Arijit, Rajat, my flatmates and others who have been very much encouraging and helped me in overcoming my loneliness.

My gratitude towards my parents is very difficult to express in words. Both my parents and my sister have given me their support since I was a little kid. They have always encouraged me to pursue my scientific interests and stood by

my side to help me in all possible ways. Even in the last stage of breast cancer and trauma associated with the three phases of chemotherapies and their adverse side effects, my mother stands by my side to support me emotionally in pursuing this doctoral degree by research. The supports from my parents, sister and other family members have been incredible, without which I couldn't have completed this thesis.

I dedicate my thesis to my parents, my sister and my other family members for their love and blessings.

Lastly, my sincere thanks to Electrical and Electronic Engineering, Nanyang Technological University, Singapore, for providing me with the full research scholarship award, which has helped me in pursuing the higher degree by research, i.e. Doctor of Philosophy.

Abstract

The ability to recognize human faces is a demonstration of incredible human intelligence. Over the last two decades researchers from diverse areas are making attempts to replicate this outstanding visual perception of human beings in machine recognition of faces. Within the face recognition literature, debate has been centered on how human beings perceive human faces and this has become an important and active research area. Psychologists concluded that holistic and component based approaches are dual routes to the face recognition. Recent studies (like FERET competition) show that holistic approaches have dominated the face recognition systems and have shown better performance than component based approaches. Although these holistic/appearance based approaches have attained certain level of maturity their performances are far away from the abilities of human recognition of faces. Owing to the immense potentiality of the face recognition applications it is imperative to develop a face recognition system, which is robust, efficient and able to achieve high recognition accuracy on large face image databases. In this thesis, we propose various algorithms which are based on statistical pattern recognition and computer vision for robust face recognition with high accuracy.

We study numerous holistic/appearance based face recognition methods among which, PCA, FDA, Bayesian maximum likelihood, NDA and their variations are the popular ones. For applications involving high dimensional patterns such as human faces, demand large number of training samples to obtain good generalization and efficient classification ability. In this thesis, we propose enhanced Bayesian face

recognition, which uses an ‘upper bound’ eigenvalue and achieves better recognition accuracy than the existing method. An eigenfeature regularization and extraction methodology is proposed in this thesis. This proposed algorithm performs regularization of the unreliable portion of the eigenspectrum using model eigenvalues. These model eigenvalues are generated using the principal/reliable portion of the eigenspectrum. Extensive experimental results and analysis on numerous databases demonstrate that this approach can consistently outperform all other popular face recognition methods.

Kernel eigenspace is decomposed using an eigenratio scheme, which has facilitated in designing a weighting/scaling function. This approach performs discriminant analysis in the whole kernel eigenspace. Experimental results on many databases demonstrate its superiority over other kernel based approaches. Lastly, the intrinsic differences between the face identification and verification are studied and a new regularization scheme is proposed for face verification purpose. This method performs subspace decomposition using two control points and regularizes some portions of the eigenspectrum using modeled eigenvalues. It consistently outperforms all other methods for face verification task. Experimental results and analysis on numerous popular face image databases demonstrate the superiority of our proposed approaches over the existing methods.

Table of Contents

Acknowledgments	i
Abstract	iii
List of Figures	x
List of Tables	xiv
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	3
1.3 Major Contributions of the Thesis	5
1.4 Thesis Organization	6
2 Literature Review	7
2.1 General Overview and Discussions on Face Recognition	7
2.2 Literature Review - Face Recognition from Intensity Images	9
2.2.1 Holistic/Appearance Based Approaches	10
2.2.2 Feature/Component Based Approaches	18

TABLE OF CONTENTS

vi

2.2.3	Hybrid Approaches	23
3	Improved Bayesian Face Recognition	26
3.1	Introduction	26
3.2	Background of Bayesian Face Recognition	27
3.3	Shortcomings of Bayesian Face Recognition	29
3.4	Proposed Method for Eigenvalue Selection in Residual Space	32
3.5	Experimental Setup and Results	35
3.6	Summary	43
4	Eigenfeature Regularization and Extraction for Face Recognition in Linear Space	45
4.1	Introduction	45
4.2	Eigenspectrum Modeling and Subspace Decomposition	49
4.2.1	Eigenspectrum Modeling	50
4.2.2	Problems of Feature Scaling and Extraction	51
4.2.3	Subspace Decomposition	52
4.3	Eigenfeature Regularization and Extraction	55
4.3.1	Eigenspectrum Regularization	56
4.3.2	Eigenfeature Extraction	59
4.3.3	The Proposed Algorithm	60
4.4	Analysis and Comparison	61
4.4.1	Discriminant Evaluation in a Subspace, Complementary Subspaces and the Whole Space	62

TABLE OF CONTENTS

vii

4.4.2	Eigenspectrum Regularization	63
4.4.3	Computational Complexity	65
4.5	Experiments	66
4.5.1	Results on FERET Database 1	66
4.5.2	Results on FERET Database 2	69
4.5.3	Results on ORL Database	70
4.5.4	Results on AR Database	72
4.5.5	Results on Georgia Tech Database	74
4.5.6	Significance of the Proposed Approach	75
4.6	Summary	76
5	Complete Discriminant Evaluation and Feature Extraction for Face Recognition in Kernel Space	78
5.1	Introduction	78
5.2	Kernel Feature Scaling and Subspace Decomposition	81
5.2.1	Overview of Kernel Discriminant Analysis	81
5.2.2	Problems in Feature Scaling and Extraction of KFDA	83
5.2.3	Eigenratiospectrum and Subspace Decomposition	84
5.3	Scaling of Kernel Eigenvectors and Feature Extraction	86
5.3.1	Scaling of Kernel Eigenvectors	86
5.3.2	Kernel Eigenfeature Extraction	89
5.3.3	The Proposed Algorithm	90
5.3.4	Computational Complexity	91

TABLE OF CONTENTS

viii

5.4	Experiments and Discussions	92
5.4.1	Results on FERET Database 1	93
5.4.2	Results on FERET Database 2	96
5.4.3	Results on ORL Database	97
5.4.4	Results on Georgia Tech Database	98
5.4.5	Summary of Experiments	99
5.4.6	Contributions of the Proposed Approach	99
5.5	Summary	100
6	Face Verification Using Eigenmodel	101
6.1	Introduction	101
6.2	Proposed Algorithm	103
6.3	Analysis and Comparisons	106
6.4	Experimental Results	107
6.4.1	Results on AR Database	108
6.4.2	Results on FERET Database 1	110
6.4.3	Results on FERET Database 2	112
6.5	Summary	114
7	Conclusions and Future Work	116
7.1	Conclusions	116
7.2	Future Work	118
	Bibliography	120

TABLE OF CONTENTS	ix
Publications	134

List of Figures

2.1	A face recognition system.	8
2.2	A Principal Component Decomposition.	12
2.3	Difference in projections between PCA and FDA.	13
2.4	Real part of Gabor kernels at 5 scales and 8 orientations.	21
2.5	Fiducial points and grids for different poses (left); Bunch Graph (right) [1]. (Courtesy of L. Wiskott, J. Fellous, and C. Malsburg). . .	21
3.1	(a) Decomposition of $\mathbb{R}^n$ into the principal subspace F and its orthogonal compliment $\bar{F}$ for a Gaussian density., (b) a typical eigenvalue spectrum and its division into two orthogonal subspaces. (Courtesy of B. Moghaddam, T. Jebera and A. Pentland [2]).	30
3.2	Weighting functions of (3.6), (3.7) and (3.8) in principal and complementary subspaces based on a typical real eigenspectrum.	33
3.3	Sample images from the FERET database for four subjects with two sample images per subject, mask and their normalized versions.	36
3.4	Recognition rate against the number of features used in the matching on the FERET database of 994 training images (497 subjects) and 1394 testing images (697 subjects).	38

3.5	Recognition rate against the rank plots for various number of feature vectors (a) $m=200$, (b) $m=150$, (c) $m=100$, (d) $m=50$ used in the matching of FERET database of 1400 training images (700 subjects) and 988 testing images (494 subjects).	40
3.6	Recognition rate against the rank plots for various number of feature vectors (a) $m=200$, (b) $m=150$, (c) $m=100$, (d) $m=50$ used in the matching of FERET database of 994 training images (497 subjects) and 1394 testing images (697 subjects).	41
3.7	Recognition rate against the rank plots for various number of feature vectors (a) $m=200$, (b) $m=150$, (c) $m=100$, (d) $m=50$ used in the matching of FERET database of 500 training images (250 subjects) and 1888 testing images (944 subjects).	42
4.1	A typical real eigenspectrum, its model and feature weighting/extraction of PCA (left); the first 50 real eigenvalues and the values of the model (right).	51
4.2	A typical real eigenspectrum, its model sorted in descending order of the face component and their inverse; decomposition of the eigenspace into face-, noise- and null-subspaces.	53
4.3	Weighting functions of (4.10) and (4.14) in the face-, noise- and null-subspaces based on a typical real eigenspectrum.	57
4.4	Sample images from the FERET database for four subjects with two sample images per subject.	67
4.5	Recognition error rate against the number of features used in the matching on the FERET database of 500 training images (250 people) and 1888 testing images (944 people).	67

LIST OF FIGURES

xii

4.6	Recognition error rate against the number of features used in the matching on the FERET database of 512 training images (128 people) and 512 testing images (128 people).	69
4.7	Sample images from the ORL database for two subjects with eight sample images per subject.	70
4.8	Recognition error rate against the number of features used in the matching on the ORL database of 200 training images (40 people) and 200 testing images (40 people).	71
4.9	Sample images from the AR database for one subject with fourteen sample images.	73
4.10	Recognition error rate against the number of features used in the matching on the AR database of 525 training images (75 people) and 525 testing images (75 people).	73
4.11	Sample images from the GT database for two subjects with eight sample images per subject.	74
5.1	Eigenratiospectrum (5.13) from a typical real kernel eigenspectrum of $\mathbf{S}_\Theta^w$	85
5.2	Weighting functions of (5.12) and (5.16) in the principal- and unreliable-subspaces based on a typical real kernel eigenspectrum.	88
5.3	Recognition error rate against the number of features used in the matching on the FERET database of 500 training images (250 subjects) and 1888 testing images (944 subjects).	93
5.4	Recognition error rate against the number of features used in the matching on the FERET database of 500 training images (250 subjects) and 1888 testing images (944 subjects).	95

LIST OF FIGURES

xiii

5.5	Recognition error rate against the number of features used in the matching on the FERET database of 994 training images (497 subjects) and 1394 testing images (697 subjects).	95
5.6	Average recognition error rate against the number of features used in the matching on the FERET database of 512 training images (128 subjects) and 512 testing images (128 subjects).	96
5.7	Recognition error rate against the number of features used in the matching on the ORL database of 200 training images (40 subjects) and 200 testing images (40 subjects).	97
6.1	Weighting functions of (4.10) and (6.2) in the principal- and unreliable-subspaces based on a typical real eigenspectrum.	104
6.2	False rejection rate against the false acceptance rate on the AR database of 420 training images (60 subjects), 420 probe genuine images (60 subjects) and 210 probe imposter images (15 subjects). The total number of genuine matches is $7 \times 60 \times 7 = 2,940$ and the total number of imposter matches is $15 \times 14 \times 60 \times 7 = 88,200$. . .	109
6.3	False rejection rate against the false acceptance rate on the FERET database of 994 training images (497 subjects), 697 probe genuine images (697 subjects) and 697 probe imposter images (697 subjects). The total number of genuine matches is $697 \times 1 = 697$ and the total number of imposter matches is $697 \times 696 = 4,85,112$	111
6.4	False rejection rate against the false acceptance rate on the FERET database of 600 training images (200 subjects), 200 probe genuine images (200 subjects) and 224 probe imposter images (56 subjects). The total number of genuine matches is $200 \times 3 = 600$ and the total number of imposter matches is $56 \times 4 \times 200 \times 3 = 1,34,400$	113

List of Tables

3.1	Recognition Rate (%) of Different Approaches on Different Data Sizes	43
4.1	Recognition Error Rate of Different Approaches for Different Number of Features.	68
4.2	Recognition Error Rate of Different Approaches for Different Number of Features.	75
5.1	Recognition Error Rate of Different Approaches for Different Number of Features.	98
6.1	Equal Error Rate (ERR %) of various approaches for three different databases.	113

Chapter 1

Introduction

1.1 Motivation

After the September 11, 2001 incident on the World Trade Center and the Pentagon followed by London bombings on July 7, 2005 there is an ever-growing need to identify and verify individuals automatically. The current state-of-art technologies of using a personal identification number (PIN) or password for these purposes are inadequate because they are disclosable, transferable and difficult to manage/remember. **Biometrics** based identification and verification methods are emerging as highly reliable and user-friendly. They can be defined as measurable physiological and/or behavioral characteristics that can be used to authenticate/verify the identity of an individual. It has a wide range of commercial, security and law enforcement applications, which include forensic identification, access control, border surveillance and human computer interface. In addition, with the rapid progress in Internet and electronic commerce, there is also a growing need to authenticate the identity of a person for secure transaction processing.

The idea of biometrics is said to date back to ancient Egypt, when records of distinguishing features and bodily measurements were used to make sure that people were who they claimed to be. Modern computer based biometric systems or so

called machine recognition systems are employed for three basic functions. The first is identification (“who is this person?”), in which a subjects identity is determined by comparing a measured biometric against a database of stored records a one to many comparisons. The second is verification (“is this person who they claim to be?”), which makes a one to one comparison between a measured biometric and one known to come from a particular person. The third is biometric watch list, which is comparatively new idea of implementing watch list cameras. Images captured by this watch list cameras can be automatically compared against criminal or other agency databases. If a match is found, then alerts can be sent out from the comprehensive surveillance solution through the common security interface. Essentially this watch list is a combination of both identification and verification processes. These processes are accomplished by using computer technology in a non-invasive way to match patterns of live individuals in real time against enrolled records. Examples include that recognize faces, hands, fingerprints, signatures, irises, voices, etc.

Among all the various biometric systems like fingerprint, hand, iris, voice, etc, face recognition can be an important and promising alternative for selecting and developing optimal biometric system. Its main advantages are that it does not require physical contact with image capturing device (camera) or non-intrusive, unobtrusive and discrete. The person can be recognized at a distance without having any prior knowledge that he/she is being identified or verified. Face recognition system does not require any advance hardware. The infrastructure for its implementation is inexpensive and already widespread. It can be used with existing image capture devices (web cameras, security cameras, etc). Also, we human beings recognize each other just by looking at each other from a distance. Even after several years we can recognize individuals just by looking into his/her 2D pictures. Hence, face recognition is suitable for many applications, taking into account its convenience for the user. Subspace based face recognition is gaining popularity because it does not require to (automatically) separate various face components [3]. In general, these methods use various statistical pattern recognition

procedures to capture the cues that are useful for discrimination. In this thesis, we explore various subspace face recognition methods for biometric identification and verification. Various methods developed for face biometric are discussed, evaluated and compared against the existing ones on large face databases.

1.2 Objectives

After 30 years of intensive research, feasible face recognition solutions are available in the biometrics market. However, the current state-of-art face recognition technology lacks maturity and fails pathetically for large-scale deployment in terms of recognition accuracy, scalability, ease of usage, computation efficiency, robustness and various other factors. According to the survey by “Biometric Technology Today” the biometric technologies market share for face recognition is still very low as compared to other technologies like hands, fingerprints, signatures, irises or voices, etc. The major obstacles faced by the state-of-art methodologies include: large intra-class variability and huge inter-class similarity; segmentation of faces from outdoor/crowd environment; noisy input and population coverage; template aging and update. Nevertheless, face recognition research has received considerable attention in recent years (and it is expected to continue) among computer vision, pattern recognition, image processing society and many others. Depending on the nature of the application, the sizes of the training and testing databases, clutter and variability of the background, noise, occlusion and finally speed requirements, some of the subtasks can be very challenging. In this dissertation, we focus on the subtasks of face recognition which include identification and verification. In identification problems, the input to the system is an unknown face (called probe) and the system output is the identity from a database (called gallery) of known individuals. On the other hand, in verification problems, the system needs to confirm or reject the claimed identity of the input face. With these things in mind we set ourselves the task of face recognition from still images, the general problem

statement can be formulated as follows: given still images of a scene, identify or verify one or more persons in the scene using the stored database of faces.

Subspace discriminant analysis is a general framework using which we can construct good classifiers (both linear and nonlinear or kernel) [4, 5] and can avoid the difficulty of the (automatic) separation of various components (like eyes, nose, mouth, etc) in face image. However, it requires many good representative sample images per person for good generalization during training. Much of the work in face recognition literature has concentrated on improving the performance of basic statistical methods used for discriminant analysis. Researchers have been using numerous discriminant analysis, sometimes along with the help of principal component analysis for improving the classification of face images and some progress have been achieved. However, due to the high dimensionality of face image, limited number of training samples and noises, the construction of a good classifier is a very difficult task. In the existing literature there are many popular methods which attempt to solve these problems. They and their variations have received considerable attention in recent times [3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20].

In this thesis, we propose robust methods to improve the current state-of-art face recognition accuracy including face identification and verification processes. Discriminant evaluation in a subspace, complementary subspaces and the whole space can lead to very different recognition results. This discriminative assessment plays an important role in how to select the features which are effective for classification purposes. The scaling or weighting of various eigenfeatures are crucial for successive discriminant evaluation. How to extract stable and discriminative features which are effective for face recognition using limited number of training samples is a challenging problem. Keeping these things in mind, we develop useful face recognition methods, which facilitate a discriminative and stable/robust low-dimensional feature representation of the face images effective for identification and verification purposes.

1.3 Major Contributions of the Thesis

The main contributions in this thesis are presented in chapters 3, 4, 5 and 6. Chapter 2 gives an introduction and the current state-of-art face recognition literature review. Chapter 7 concludes this thesis. The most important individual contributions are:

- Bayesian face recognition is reported to be one of the best performer in face recognition. We have proposed an enhancement to this approach, which selects a more robust weighting parameter and introduced the term - ‘upper bound’ in place of the existing average eigenvalue over the residual space. Our proposed approach has outperformed the existing method when experimented on Face Recognition Technology (FERET) face image database (chapter 3).
- A novel approach called eigenfeature regularization and extraction for face recognition in linear space is proposed. Face eigenspace of the within-class scatter matrix is decomposed into three subspaces. Eigenfeatures are regularized differently in these three subspaces based on an eigenspectrum model. The proposed algorithm outperforms all other popular subspace face recognition methods (chapter 4).
- A kernel-machine based algorithm, which utilizes the ratios of the eigenvalues is proposed. This algorithm circumvents undue scaling of the eigenvectors corresponding to the unreliable small and zero eigenvalues and extracts features by the discriminant evaluation in the whole kernel space. It outperforms all other popular kernel based face recognition methods (chapter 5).
- A face verification method based on the partial modeled eigenvalues and eigenratiospectrum is proposed. This algorithm can handle face verification better than all other popular methods (chapter 6).

1.4 Thesis Organization

This thesis is organized as follows: In Chapter 1, motivations, objectives and the major contributions of this thesis are presented. Chapter 2 provides a general overview and discussions on face recognition systems. A detail literature review is given in this chapter. In chapter 3, we propose an enhancement to the Bayesian face recognition algorithm. Eigenfeature regularization and extraction in linear space for face recognition is presented in chapter 4. Chapter 5 deals with our proposed approach on complete discriminant evaluation and feature extraction in kernel space. Face verification using a partial modeled eigenspectrum and eigenratios is presented in chapter 6. Finally, conclusions and suggestions for future work are presented in chapter 7.

Chapter 2

Literature Review

2.1 General Overview and Discussions on Face Recognition

Among the thousands of objects that humans are able to recognize, faces are arguably the most ubiquitous and the most ecologically important - so much so that in order to be a competent member of any social group, all people must become face experts. As a mother studies the face of her newborn infant, so too does the infant explore the face of its mother. The newborn rapidly learns to discriminate its mother's face from that of a stranger at a tender age of 45 hours [21]. By adulthood, it will be able to discriminate and recognize hundreds or even thousands of faces. This ability is particularly impressive given that all faces share the same parts in a common arrangement. Research with animals has also shown that these capabilities are not unique to humans. Sheep, for example, are known to have a remarkable memory for faces and can remember and identify other sheep even after years of separation [22]. How the mammalian brain solves the problem of visual face recognition has been a topic of study since the early days of brain science. Psychological studies among humans show that holistic processing and components processing represent two routes to face recognition and the processed

features vectors play a critical role in face recognition. It can be assumed that the representation of any face pattern lies along a holistic-to-component continuum where some features are weighted more toward holistic representation and others are biased toward component based representation [23].

In this study, we configure a general face recognition system which consists of four processing parts. A general system overview is shown in Fig. 2.1. The input

Figure 2.1: A face recognition system.

to the system can be either image or video and then subsequently the following steps are performed.

- (a) *Face detection/segmentation or tracking*: The detection of faces from an arbitrary uncontrived image is a critical precursor and has a major influence on the performance and usability of the whole system. Robust face detection algorithm is needed to determine face location as well as its precise placement with the help of eye coordinates and/or nose or other fiducial points. This detection procedure must operate flexibly and reliably regardless of illumination/lighting conditions, background clutter in the image, multiple faces in the image, as well as variations in face position, scale, occlusions, expression and pose. Face detection is absolutely necessary for achieving good recognition rate because faces are a special kind of objects, which are further processed as shown in Fig. 2.1. If they are detected wrongly then the subsequent processes may fail or the system will achieve very low recognition accuracy.

- (b) *Face alignment*: After the successful detection of face, it would be beneficial

to align and/or rotate the face so that the eye coordinates are placed at a specific distance on the same horizontal axis. In order to achieve these, scaling and/or rotation of face would be required, thereby the input face image is normalized in facial geometry. For subspace face recognition usually a mask is applied to remove all the hair, neck and ear portions. Finally histogram equalization is applied to the masked image for photo normalization.

- (c) *Feature extraction*: Training occurs in this step, which uses various exemplars available from the training data set and extracts low dimensional feature vectors. This is achieved by applying various face recognition algorithms which explore attributes useful for distinguishing different persons. Low dimensional effective discriminating features of the enrolled subjects are extracted and stored in the database.
- (d) *Feature matching*: This step is the testing/matching in which the stored features in the database are compared with the probe/test image using similarity or distance measures. After the feature matching, a similarity score or distance measure is generated. The highest similarity score or minimum distance measure provides the identity of a person. On the contrary, a person is verified as genuine or imposter if the similarity score or distance measure is above or below the specified threshold respectively.

2.2 Literature Review - Face Recognition from Intensity Images

One of the successful applications of computer vision, pattern recognition, image analysis and understanding is face recognition which has recently received significant attention, especially during the last two decades. There are several reasons for this trend. Most importantly, it has a wide range of commercial and law enforcement applications, which include forensic identification, access control and human

computer interface. Although the machine recognition of faces has attained certain level of maturity, their success has been thwarted by the constraints imposed by many real applications. For example, face recognition which deals with diverse change in pose along with illumination and expression changes in an outdoor environment is still an ongoing research area. In other words, the current face recognition systems are still far away from the capability of the human recognition/perception system. In this section we briefly review the existing literature and some insights into the studies of machine recognition of faces [3, 24, 6].

Numerous face recognition methods have been proposed during the past 30 years. Face recognition is such a challenging yet interesting problem that it has attracted researchers who have different backgrounds: computer vision, neuroscience, pattern recognition, machine learning, image processing, neural networks, psychology and computer graphics. It is due to this fact that the literature on face recognition is vast and diverse. Broadly we can classify the face recognition approaches into three categories:

2.2.1 Holistic/Appearance Based Approaches

Nearly a century ago, psychologists characterized the holistic/appearance based visual apprehension as a perceptual gestalt where the perception of the whole stimulus takes precedence over the sum of its individual parts. Face recognition provides the prime example of gestalt perception, this concept has proved to be elusive and difficult to operationalize in the laboratory [23]. In holistic approach, the features of a face (nose, eyes, mouth, cheek, other parts) are not independently specified in face representation. According to this hypothesis, the primary representation in face recognition is at the level of the whole face. Hence, faces are recognized in terms of their undifferentiated wholes and not in terms of their constituent parts. In face recognition, the holistic hypothesis predicts that the recognizability of the part is significantly enhanced when it is embedded in the whole face than when shown in isolation. Psychologists have experimentally defined “holistic processing

as the improved ability to recognize a part stimulus when presented in the context of the whole stimulus than when presented in isolation” [23]. Using this criterion, it has been shown that faces are recognized more holistically than other kinds of objects (e.g. houses, inverted faces, scrambled faces). Below we study some of the popular holistic/appearance approaches:

- (a) *Principal Component Analysis*: In machine holistic/appearance recognition of human faces the whole face image is input to the system. Bigger the face image more the information it will contain and hence, more processing time it would require. However, modern statistical tools can help in reducing this processing time. For example, principal component analysis (PCA) is a statistical unsupervised dimensionality reduction technique, which can be used to find the principal projections/directions of the data with larger variances. Fig 2.2 shows how the PCA method finds the optimal method of representing data. The two clusters of data can be represented either along axis 1 or axis 2 but we can see that if this data is represented along axis 1 they are well separated, the objective of PCA is to represent data in most compact form or in the least sum of squared error sense. Actually the magnitude of the axis (1 or 2) is of no real significance, it is the direction which determines the best compact representation of the data in reduced dimension.

Kirby and Sirovich [25, 26] first applied this method, which is based on Karhunen-Loeve transformation (KLT) for the characterization of human faces and later extended by Turk and Pentland [27]. It produces an expressive subspace for face representation that is spanned by a set of eigenvectors corresponding to the leading eigenvalues of the covariance matrix of the training samples. The first advantage of PCA is that it can represent data in the most compact form from its high dimensional space in a minimum squared error sense such that it has most optimal reconstruction error. The error in its reconstruction is the summation of the eigenvalues corresponding to the discarded eigenvectors [4]. Secondly, the model parameters can be computed

Figure 2.2: A Principal Component Decomposition.

directly from the training data, for example, by diagonalizing the sample covariance matrix. There are several other attractive features of PCA, like it is less sensitive to the training data, number of samples per subject and noises present in the data. All these striking features of PCA made it the most popular statistical dimensionality reduction tool. Due to its popularity, it has been extended to nonlinear or kernel based subspace methods called kernel principal component analysis (KPCA) [28, 6, 29].

- (b) *Fisher Discriminant Analysis*: Despite of all the attractive features of PCA, it is not a popular tool for classification because it does not encode any class information. For classification, Fisher discriminant analysis (FDA) [30, 4, 31, 32] (also known as linear discriminant analysis LDA) is superior to PCA, when enough training samples per subject are available [33]. FDA is a supervised technique which encodes class information and searches for feature basis vectors for which the variance of between-class scatter matrix is maximum and variance of the within-class scatter matrix is minimum. Fig. 2.3 shows the difference between PCA and FDA objectives. It shows two clusters of data belonging to class 1 and class 2. Comparing the two projections in the Fig. 2.3, PCA actually smears the classes together so that

Figure 2.3: Difference in projections between PCA and FDA.

they are no longer linearly separable in the projected space. It is clear that although PCA achieves larger total scatter, FDA achieves greater between-class scatter and consequently, classification is simplified.

However in practical applications, it is very difficult to achieve the Fisher objective criteria because of the limited number of training samples and noises present in the data [4]. More specifically, singularity problem can occur to any of the scatter matrices, which hinders the successful application of Fisher criteria. Swets and Weng [31] proposed a two stage PCA followed by FDA method, also known as the Fisherface method [32], in which PCA is first applied to original image data for dimensionality reduction so as to make within-class scatter matrix nonsingular before the application of FDA. However, applying PCA for dimensionality reduction may lose important discriminative information [14, 34, 35, 36, 37, 38]. Rank decomposition method for FDA is proposed in [39], which is based on successive eigenvalue decomposition of the total scatter matrix and between-class scatter matrix. Perturbation method is used in [40], where a small perturbation matrix is added to the within-class matrix in order to make it nonsingular. Fractional-

step linear discriminant analysis is proposed in [41]. This method performs dimensionality reduction in a few small fractional steps allowing for the relevant distances to be more accurately weighted. Although this method has been successfully tested on low-dimensional patterns whose dimensionality is less than 5, it cannot be directly applied to high-dimensionality patterns, such as those face images. FDA is theoretically sound and promises to find the most optimal directions suitable for classifications. Hence, they are extended to nonlinear or kernel based approaches. For example, kernel Fisher discriminant analysis is proposed in [42]. Additionally, there are many recently proposed methods available in the literature [3, 6, 7]. We shall study them in detail and also their short comings in the introduction of chapters 4 and 5.

- (c) *Evolutionary Pursuit*: An evolutionary pursuit (EP) based adaptive representation and its application to face recognition was presented in [43]. EP learns to form the optimal basis for data compressions as well as pattern classification. It implements strategic characteristics of genetic algorithms for finding the space of possible solutions to determine the optimal basis. EP processes face images in a lower dimensional whitened PCA subspace. Directed random rotations of the basis vectors in this space are then searched by genetic algorithms where evolution is driven by a fitness function defined in terms of performance accuracy (empirical risk) and class separation (confidence interval). The viability of this method has been shown for face recognition, where the large number of possible bases requires a greedy search algorithm [43]. The authors reported better performance than PCA or eigen-faces [27] and better generalization capability than Fisherfaces [32].
- (d) *Bayesian Face Recognition*: Another popular method called Bayesian face recognition is proposed in [2]. (We call maximum likelihood for face recognition [2] as ML for a symbolic representation). Its subspace is spanned by the difference of two face images. Such difference images fall into two dis-

tinct classes, intrapersonal and extrapersonal and form distributions that are approximately Gaussian [2, 44]. A large performance improvement of this Bayesian probabilistic matching technique over standard nearest-neighbor eigenspace matching was reported using large FERET face data sets in [45].

- (e) *Independent Component Analysis*: Independent component analysis (ICA) [46, 5] has been applied to face recognition successfully [47, 48, 49, 50]. The concept of ICA may actually be seen as an extension of the PCA. Unlike PCA, it seeks directions that are most independent from each other and is also related to the blind source separation [5]. ICA searches for a transformation that minimizes the statistical dependence between its components [51, 52]. Second-order statistics capture the amplitude spectrum of images but not their phase spectrum. The higher-order statistics capture the phase spectrum. In a task such as face recognition, much of the important information may be contained in the higher-order relationships among the image pixels. Thus, generalizations of PCA which are sensitive to higher-order relationships are important, ICA is one such generalization.

Bartlett *et al.* [47] used a version of ICA derived from the principle of optimal information transfer (called infomax [53, 54]) through sigmoidal neurons. The goal of the infomax principle is to maximize the mutual information between the environmental variable and the output of the neural network [54]. Face recognition using ICA is performed in [47] on FERET database using two different architectures, one which treated the images as random variables and the pixels as outcomes, and a second which treated the pixels as random variables and the images as outcomes [47]. The metric induced by ICA can be superior to PCA in the sense that it may provide a representation more robust to the effect of noise. It is possible for ICA to be better than PCA for reconstruction in noisy or limited precision environments [55, 47]. ICA has also been extended to kernel ICA [56]. However, there are many disadvantages of using ICA over other de-correlations methods. For example,

there may be no ICA solution and also, a given ICA algorithm may not find the solution even if it exists, since there are approximations involved [57]. Unlike PCA, there is no definite method of how to select the features or basis vectors. In practice, the different formulations of the independence constraint are designed to enable different approximation techniques, and the algorithms find different solutions because of differences among these techniques. Moreover, there are recent studies which show that PCA can outperform ICA [58, 59].

- (f) *Tensorfaces*: Multi-linear based tensorfaces was proposed in [60, 61]. A tensor is a multidimensional generalization of a matrix. Multi-dimensional tensor are decomposed by multi-mode singular value decomposition (SVD). Their core tensor governs the interactions between mode matrices, which contain the orthonormal bases for the spaces spanned by the corresponding dimensions of the data tensor. The mode matrices can be obtained by flattening the tensor across the corresponding dimension and performing PCA on the columns of the resulting matrix. They considered tensor of order 5 taking into consideration of face images with viewpoints, illuminations and expressions [60]. The authors reported the performance of tensorfaces to be significantly better than that of the standard eigenfaces [27].
- (g) *Neural Network*: Lin *et al.* [62] proposed probabilistic decision based neural network (PDBNN) for face recognition and detection. This is based on decision-based neural networks for signal/image classifications in [63]. PDBNN is a fully automatic face detection/recognition system, in which only eye-brows, eyes and nose regions are selected and processed to produce two features at reduced resolutions. These features are fed into two PDBNNs and the final recognition result is the fusion of the outputs of these two PDBNNs. PDBNN has a lower false acceptance/rejection rate as compared to most multi-class recognition systems that use a discrimination function between any two classes. The architecture is also good for hardware imple-

mentations. However, it is unclear how to estimate the full density functions for the classes with small number of samples. Furthermore, the system can have problems when number of classes increases.

- (h) *Support Vector Machines*: One last subspace face recognition method that we review in this category is by using support vector machines (SVM) [64]. After Phillips [65] introduced SVM for face recognition, many new different methods of SVM have been presented for face identification or verification [66, 67, 68]. In this theory, given a set of points belonging to two classes, a SVM finds the hyperplane that separates the largest possible fraction of points of the same class on the same side, while maximizing the distance from either class to the hyperplane. Vapnik [64] called this hyperplane as optimal separating hyperplane (OSH) which minimizes the risk of misclassifying not only the examples in the training set but also the unseen examples of the test set. Features are extracted using PCA and then the discrimination functions between each pair are learned by SVMs. Subsequently, the disjoint test set enters the system for recognition. In [67], SVM is compared with the standard eigenface approach [27] using nearest center classification (NCC) criterion for face recognition and the former was found to perform better. However, SVMs can get over-trained and may result in the loss of generality. They involve many parameters and can employ different kernels. How to choose these parameters and kernel functions are open questions in the literature. This makes the optimization space rather extensive, without the guarantee that it has been fully explored to find the best solution.

In general, the present state-of-art holistic/appearance based face recognition approaches use subspace based methodologies. Recently, these subspace based methods have gained much popularity and have shown to perform better than the component based approaches [3, 6].

2.2.2 Feature/Component Based Approaches

What mechanisms underlie the successful discrimination of so many faces? As with other classes of visual stimuli, the individual components or features of a face can serve as one source of information (e.g. the size of the nose or eyes). However, component information may be of limited use given the common configuration and similarity across faces. There is considerable evidence that the spatial relations among the components of the face (e.g. the relative position of the nose and mouth), and/or its global configuration or gestalt, play a more important role in face recognition [23]. Unlike the above appearance/holistic based methods the preprocessing of these methods includes identification of various components like eyes, nose, mouth, etc in the face images. It is still a challenging job to find these components in the face image automatically. In the existing literature there are very few methods which can detect these components automatically. It is also sometimes required to fix some fiducial points (points which do not change much due to change in expression, light or pose variations, e.g. eye coordinates, nose tip, etc) in the input images, this is done manually in the training phase but as the training matures the systems can identify these points automatically with less and less error. Below we discuss some of the well known methods:

- (a) *Dynamic Link Architecture*: Dynamic link architecture (DLA) [69] is a component based approach which attempts to solve some of the conventional problems of artificial neural networks (ANN). One of the most important of these being the representation of syntactical relationships in NN. Neural sheets of hypercolumns which are composed of simple feature detectors are modeled as Gabor wavelets. Each object is represented in memory model sheet which is a two-dimensional array of features. The match of the image to the models is performed by network self-organization in which rapid reversible synaptic plasticity of the connections (“dynamic links”) between the two domains is controlled by signal correlations, which are shaped by fixed inter-columnar connections and by the dynamic links themselves. A typical

local feature representation consists of wavelet coefficients based on Gabor transform [70, 71] for different scales and orientations based on fixed bases called jets [69]. The connectivity matrices are initialized using the similarities between the jets of the connected neurons. DLA serves as a process to restructure the connectivity matrices and to find the correct mapping between the models and the image. The models cooperate with the image depending on their similarity. A simple winner-take-all mechanism sequentially rules out the least active and least similar models and the best fitting one eventually survives [69].

- (b) *Elastic Bunch Graph Matching*: One of the most successful among all the feature/component based approaches is the elastic bunch graph matching algorithm (EBGM) [1]. This algorithm is an extension of dynamic link architecture [69]. Unlike the above DLA method where only a single jet is attached to each node, authors of EBGM algorithm attached a set of jets called the bunch graphs each derived from a different face image. These locally estimated wavelet coefficients are robust to illumination change, translation, distortion, rotation and scaling. The algorithm involves a normalization and preprocessing step, details of which can be found in [72]. EBGM approach is summarized in the following steps:

- (i) Preprocessing with Gabor wavelets: The Gabor wavelets are biologically motivated convolution kernels in the shape of plane waves restricted by a Gaussian envelope function [1, 72]. The set of convolution coefficients for kernels of different orientations and frequencies at one image pixel is called a jet.

Jets: A jet describes a small patch of grey values in an image X_{ij} , which is the image j of person i around a given pixel location (x, y) . It is based on a wavelet transform, defined as a convolution with the kernel

$$G(x, y, \theta, \varphi, \nu, \varrho) = \exp^{-(x'^2 + \varrho^2 y'^2)/2\nu^2} \cos(2\pi x'/\chi + \varphi),$$

where $x' = x \cos \theta + y \sin \theta$ and $y' = -x \sin \theta + y \cos \theta$.

θ specifies the orientation of the wavelet. This particular set uses eight different orientations over the interval from 0 to π . i.e. $\theta \in \{0, \pi/8, 2\pi/8, 3\pi/8, 4\pi/8, 5\pi/8, 6\pi/8, 7\pi/8\}$.

χ specifies the wavelength of the sine wave. This set starts at 4 pixels and continues to longer wave length at half octave intervals. i.e. $\chi \in \{4, 4\sqrt{2}, 8, 8\sqrt{2}, 16\}$.

φ specifies the phase of the sine wave. Typically Gabor wavelets are either even or odd. The even form of the sine wave corresponds to a cosine function and the odd form corresponds to a sine function. i.e. $\varphi \in \{0, \pi/2\}$.

ν specifies the radius of the Gaussian. This parameter is proportional to the wavelengths, such that the wavelets of different sizes and frequencies are scaled versions of each other. i.e. $\nu = \chi$.

ϱ specifies the aspect ratio of the Gaussian. The wavelets generally used have circular Gaussian. i.e. $\varrho = 1$.

This yields 8 orientations, 5 frequencies and 2 phases for a total of 80 different wavelets. We show only 40 different wavelets corresponding to $\varphi = 0$ only. Fig 2.4 shows the 40 wavelet filters.

- (ii) Landmark localization: This is the bunch graph creation step. The edges are labeled with distance information and nodes are labeled with wavelet responses locally bundled in jets. A bunch graph is created in two stages. Its qualitative structure as a graph (a set of nodes plus edges) as well as the assignment of the corresponding labels (jets and distances) for one initial image is designer-provided, where as the bulk of the bunch graph is extracted semi-automatically from sample images by matching the embryonic bunch graph to them, less and less often intervening to correct incorrectly identified fiducial points as shown in fig. 2.5 (left). These designers provided image bunch graphs are known as model graphs and are stored in the database as shown in fig 2.5

Figure 2.4: Real part of Gabor kernels at 5 scales and 8 orientations.

(right). Stored model graphs can be matched to new images to generate image graphs, which can be incorporated into a gallery and become model graphs.

Figure 2.5: Fiducial points and grids for different poses (left); Bunch Graph (right) [1]. (Courtesy of L. Wiskott, J. Fellous, and C. Malsburg).

- (iii) Comparing jets: Measuring the similarity for two jets is fundamental and is required for both landmark localization and face graph similarity measurement. The final phase similarity measure is as given in [72].

The modulus of the coefficients of two different jets of the same fiducial points in different images obtained after convolution with the Gabor kernel.

- (iv) Recognition: After having extracted model graphs from the gallery images and image graphs from the probe images, recognition is possible with relatively little computational effort by comparing an image graph to all model graphs and picking the one with the highest similarity values. The similarity function used for comparing the graphs can be an average over the similarities between pairs of corresponding jets. A person is recognized correctly if the correct model yields the highest graph similarity measure.

This method is general, flexible and has achieved good recognition performance [73, 1]. It is designed for an in-class recognition task, i.e. for recognizing members of a known class of objects. It has the potential for specialization of the system for specific object types and to handle different poses with the help of object-adapted grids. The face bunch graph is able to represent a wide variety of faces, which allows matching on face images of previously unseen individuals. Even if the person of the new image is not included in the face bunch graphs, the image graph reliably refers to the fiducial points. This considerably accelerates recognition from large databases, since for each probe image, node positions need to be searched only once instead of each attempted match to a gallery image. The success of the EBGM system may be due to its resemblance to the human visual system [74]. However, this method is very tedious in its practical implementations as all the initial face graphs are to be created manually. Although this method is robust to pose and expression changes, its sensitivity with respect to illumination variations is yet to be tested. Furthermore, if more number of fiducial points are selected for improving accuracy then the system can be computationally complex and manually very labor intensive.

- (c) *Facial parts based recognition*: Brunelli and Poggio in [75] performed face recognition by independently matching templates of three facial regions (both eyes, nose and mouth). The configuration of the components during classification was unconstrained since the system did not include the geometrical model of the face. In [76], window was shifted over the face image and the discrete cosine transform (DCT) coefficients computed within the window were fed to a 2-D hidden Markov model (HMM). Recently, a probabilistic approach using part-based matching has been proposed in [77] for expression invariant and occlusion tolerant recognition of frontal faces. This approach is good and much robust against expression changes and occlusions.

2.2.3 Hybrid Approaches

Just as the human visual perception system uses both local features and the whole face region to recognize a face, a machine recognition system should use both. One can argue that these methods could potentially offer the better of the two types of methods. In the present existing literature there are very few methods proposed which are based both on holistic and component based approaches. In this section we discuss some of the popular methods.

- (a) *Modular Eigenfaces*: Pentland *et al.* proposed modular eigenfaces in [78], which uses both global eigenfaces and local eigenfeatures. This was an extension of their earlier work on eigenfaces [27]. They extended the idea of eigenfaces to eigenfeatures, such as eigeneyes, eigenmouth, etc. Using only a limited number of samples, the recognition accuracy as a function of the number of eigenvectors was measured for eigenfaces only and also for the combined representation. For smaller number of subspaces eigenfeatures performed better than the eigenfaces. However, when the combined set was used only small improvement was obtained. Their experiments support that feature-based approach may be useful when gross variations are present in

the input images.

- (b) *Local Feature Analysis*: Another interesting biologically motivated method called local feature analysis (LFA) was proposed by Penev and Atick in [79]. They argued that though a huge array of receptors exist in the human retina, only a small fraction of them are active, corresponding to natural objects/signals that are statistically redundant. From these sparsely distributed receptors, the brain has to discover where and what objects (what and where to look for, this is also called *focus of attention* - FOA) are in the field of view and recover their attributes. Consequently, one expects to represent the natural objects/signals in a subspace of lower dimensionality by finding a suitable parameterizations. LFA is used to extract topographical local features from the global PCA modes. Unlike PCA kernels which do not have any local information, LFA kernels at selected grids have local support. For same number of kernels, the perceptual reconstruction quality of LFA based on the optimal set of grids is better than that of PCA. LFA is claimed to be used in Visionicss commercial system FaceIt [3].
- (c) The last kind of methods that we review in this category is shape based approaches for face recognition. Cootes *et al.* proposed active shape model (ASM) in [80]. They described methods for building models which learns patterns of variability from training set. The proposed flexible models is an iterative refinement algorithm similar to active contour models (snakes) [81]. The key difference is that it can deform and fit according to the training data, which allow considerable variability but are still specific to the class of structures they represent. A new statistical based active appearance model (AAM) is proposed in [82]. Their models of appearance can match to any of a class of deformable objects (e.g., any face with any expression, rather than one persons face with a particular expression). To increase the speed of their matching process they pre-computed the derivatives of the residual of the match between the model and the target image and use them to compute the

update steps in an iterative matching algorithm [82]. Bayesian shape model (BSM) is proposed in [83]. In this model a full-face is designed to describe the shape of a face and the PCA is used to estimate the shape variance of the face model. It incorporates the advantages of both EigenSnake model [84] and the AI-Snake mode [85]. In BSM, the full-face model consists of contour points and the control points. The contour points are the landmarks of the face outline, eyebrows, eyes, nose and mouth; the control points are the feature points that are placed in the key position of the face. Every three of the control and contour points can be used to form a face unit triangle for facial image warping. The prior distribution of the of the full-face model is estimated by PCA. The individual shape model of each component contour is also built. Therefore, not only the separate shape variation of each component contour of the facial features, but can also the global shape variation be represented conveniently [83].

In the following chapters 3, 4, 5, 6 and 7, we present our proposed approaches for improving the face recognition accuracy. Numerous advantages and disadvantages are discussed in the introduction of each chapters. Our thrust is on understanding subspace face recognition algorithms, which uses appearance/holistic based approach for face recognition.

Chapter 3

Improved Bayesian Face Recognition

3.1 Introduction

In the FERET evaluation test, Bayesian face recognition has achieved a large performance improvement over other nearest-neighbor eigenspace matching methodologies [45, 3]. This probabilistic matching method is proposed by Moghaddam and Pentland [44, 2], which differs from traditional algorithms in two important ways. Firstly, this method uses a Bayesian classifier rather than the nearest-neighbor classifier. Secondly, they cast the multi-class problem of distinguishing among images of different subjects into the problem of distinguishing between intrapersonal and extrapersonal difference images [2, 86, 44, 87].

While performing the classification of difference images as intrapersonal or extrapersonal, each difference image is regarded as a point in a high-dimensional space. The space of all possible difference images is vast and very sparsely populated. Each difference image occupies a point in this space. Teixeira in [86] pointed out that, “In the real universal, a great majority of the space is unoccupied - the majority of vacant points correspond to difference images that never arise in prac-

tice. Difference images which arise in practice will form clusters, analogous to galaxies in the real universe, and on the whole, these ‘galaxies’ will take only an infinitesimally small portion of the entire space”. One important assumption in [2, 44] is that the particular difference images belonging to the intrapersonal and extrapersonal difference images come from two such ‘galaxies’ - that is, they originate from distinct and localized Gaussian distributions within the space of all possible images.

Bayesian face recognition performs an efficient probability density estimation by decomposing the input space into two mutually exclusive subspaces: the principal subspace F and its orthogonal subspace $\bar{F}$. The likelihood functions are derived from a sample training data using an efficient subspace method for density estimation of high-dimensional data [2]. Covariances only in the principal subspace are estimated for use in the Mahalanobis distance and the distance measure in its orthogonal subspace is weighted by an estimated eigenvalue over the residual space. However, this method lacks optimal estimation of eigenvalues in some portion of the principal subspace as well as over the residual space because we cannot trust fully the lower portion of the eigenspectrum (or the covariances) obtained from the limited training data. In the following section we discuss the background of Bayesian face recognition and then analyze the problems associate with this method.

3.2 Background of Bayesian Face Recognition

Given a set of properly aligned and normalized h -by- w face images, we can form a training set of column vectors $\{X_{ij}\}$, where $X_{ij} \in \mathbb{R}^{n=hw}$ is called image vector, by lexicographic ordering the pixel elements of image j of person i . Here lexicographic ordering means scanning each pixel from left to right (row wise) and put down in a single column. This makes a single facial (column) vector $\{X_{ij}\}$ of dimensionality $n \times 1$. Let the training set contain p persons and q_i sample images for person i . The

number of total training sample is $l = \sum_{i=1}^p q_i$. Moghaddam *et al.* [2] proposed a probabilistic similarity measure based on Bayesian belief that image intensities differences denoted by $\Delta = X_{ij} - X_{st}$, where $X_{ij}, X_{st} \in \mathbb{R}^n$, of two images X_{ij} and X_{st} is an n -dimensional space are the characteristics of typical variations in appearance of an individual.

The classifier proposed by Moghaddam *et al.* defines the subspace spanned by the difference between two face images. The difference image for two face images is the signed arithmetic difference between respective pixels in the source images. Using such difference images two distinct and mutually exclusive classes can be defined. Φ^I representing *intrapersonal* between multiple images of the same individual (e.g with different lighting condition, expression and pose variations), and Φ^E representing *extrapersonal* variations in matching two different individuals. It is assumed that both classes are Gaussian distributed and seek to obtain estimates of the likelihood functions $P(\Delta | \Phi^I)$ and $P(\Delta | \Phi^E)$ for a given intensity difference $\Delta = X_{ij} - X_{st}$. The similarity is then expressed in terms of probability

$$S(X_{ij}, X_{st}) = P(\Delta \in \Phi^I) = P(\Phi^I | \Delta), \quad (3.1)$$

where $P(\Phi^I | \Delta)$ is the a posterior probability given by Bayes rule, using estimates of the likelihoods $P(\Delta | \Phi^I)$ and $P(\Delta | \Phi^E)$. These likelihoods are derived from training data using an efficient subspace method for density estimation of high dimensional data.

Given these likelihoods, we can evaluate a similarity score $S(X_{ij}, X_{st})$ between a pair of images directly in terms of intrapersonal a posteriori probability as given Bayes rule

$$S(X_{ij}, X_{st}) = P(\Phi^I | \Delta) = \frac{P(\Delta | \Phi^I) \cdot P(\Phi^I)}{P(\Delta | \Phi^I) \cdot P(\Phi^I) + P(\Delta | \Phi^E) \cdot P(\Phi^E)}, \quad (3.2)$$

where the priors $P(\Phi^I)$ can be set to reflect specific operating conditions (e.g. number of test images *vs.* the size of the database) or other sources of a priori

knowledge regarding the two images being matched [2, 86, 44, 87, 59].

Bayesian formulation in (3.2) casts a standard face recognition task, essentially an p -ary classification problem for p individuals into a binary pattern classification problem with Φ^I and Φ^E . This problem is then solved using the maximum *a posteriori* (MAP) rule - two images are determined to belong to the same individual if $P(\Phi^I|\Delta) > P(\Phi^E|\Delta)$, or equivalently, if $S(X_{ij}, X_{st}) > 1/2$, [2].

An alternative probability similarity measure can be defined in simpler form using the intrapersonal likelihood alone [2],

$$S'(X_{ij}, X_{st}) = P(\Delta|\Phi^I), \quad (3.3)$$

thus leading to the maximum likelihood (ML) recognition as opposed to MAP recognition in (3.2). Experimental results in [2] shows that this simplified ML measure can be almost as effective as its MAP counterpart in most cases. So in this thesis we consider only ML measure for face recognition. In the following section we shall discuss the procedure of estimation of $P(\Delta|\Phi^I)$ and its shortcomings for face recognition.

3.3 Shortcomings of Bayesian Face Recognition

This approach uses the image intensity difference images $\Delta = X_{ij} - X_{st}$, where $\Delta \in \mathbb{R}^n$ for computing likelihood densities. We face two problems in dealing with such high dimensional data. Firstly, since $n \gg p$, we almost always lack sufficient independent training samples to compute reliable 2nd-order statistics for the likelihood densities (e.g. resultant covariance matrix is singular). Secondly, to estimate this statistics, the computational cost in evaluating the likelihoods is formidable. Furthermore, this computation would be highly inefficient since the intrinsic dimensionality or major degrees-of-freedom of Δ is likely to be significantly smaller than n .

To deal with such high dimensional data, efficient density estimation method is proposed in [2, 86, 44, 87], which divides the vector space $\mathbb{R}^n$ into two complementary subspaces using an eigenspace decomposition. This method relies on a principal components analysis (PCA) [25, 26] to form a low-dimensional estimate of the complete likelihood which can be evaluated using only the first m components, where $m \ll n$.

Figure 3.1: (a) Decomposition of $\mathbb{R}^n$ into the principal subspace F and its orthogonal complement $\bar{F}$ for a Gaussian density, (b) a typical eigenvalue spectrum and its division into two orthogonal subspaces. (Courtesy of B. Moghaddam, T. Jebara and A. Pentland [2]).

Fig. 3.1 shows an orthogonal decomposition of the vector space into two mutually exclusive subspaces: the principal subspace F containing the first m principal components and its orthogonal complement $\bar{F}$, which contains the residual of the expansion. The component in the orthogonal subspace F is the so-called “distance-from-feature-space” (DFFS), a Euclidean distance equivalent to the PCA residual error. The component of Δ which lies in the feature space $\bar{F}$ is referred to as the “distance-in-features-space” (DIFS) and is a Mahalanobis distance for Gaussian densities [2].

$P(\Delta|\Phi^I)$ can be estimated as the product of two independent marginal Gaussian densities in principal and complementary subspaces. For any difference image Δ between two images, the probability of that image belonging to class Φ^I is

computed as follows [86]:

$$P(\Delta|\Phi^I) = \left[\frac{\exp(-\frac{1}{2} \sum_{i=1}^m \frac{y_i^2}{\lambda_i})}{(2\pi)^{m/2} \prod_{i=1}^m \lambda_i} \right] \cdot \left[\frac{\exp(-\frac{1}{2\rho} \varepsilon^2(\Delta))}{(2\pi\rho)^{(n-m)/2}} \right], \quad (3.4)$$

where y_i are principal components and $\varepsilon^2(\Delta)$ is the residual reconstruction error in the partial Karhunen-Loeve expansion [2, 87]:

$$\varepsilon^2(\Delta) \equiv \|\tilde{\Delta}\|^2 - \sum_{i=1}^m y_i^2, \quad (3.5)$$

If Φ_m^I is the m dimensional linear subspace spanned by the leading eigenvectors of the intrapersonal covariance matrix then the value $\|\tilde{\Delta}\|^2$ is the square of the vector length of the difference image Δ before projection by Φ_m^I while the sum $\sum_{i=1}^m y_i^2$ is the vector length after the projection by Φ_m^I . The average eigenvalue ρ is calculated in the following way:

$$\rho = \frac{1}{n-m} \sum_{i=m+1}^n \lambda_i, \quad (3.6)$$

$P(\Delta|\Phi^E)$ can be estimated in a similar way in the extrapersonal subspace computed from $\{\Delta|\Delta \in \Phi^E\}$. It has been shown that maximum likelihood (ML) measure $P(\Delta|\Phi^I)$ alone is almost as effective as the MAP measure [2].

The left hand term of (3.4) is nothing but the Mahalanobis distance measure, also known as distance-in-feature-space (DIFS). This captures the difference of the projected eigenfeature vectors through the subspace spanned by the m leading eigenvectors. The right hand term, $\varepsilon^2(\Delta)$ is the PCA reconstruction error weighted by the inverse of the average eigenvalues between the range m and n and can be called as distance-from-feature-space (DFFS). This out-of-space variance is important and plays a critical role in the classification of the face difference images [14, 38, 87, 88]. Due to the limited training sample size, PCA cannot estimate the eigenvalues beyond the rank of intra-class variation matrix, which is $r \leq \min(n, l - p)$. As a result of which, Bayesian probabilistic approach has to

perform the minimization of a cost function with respect to ρ [44]. Although this guides to the minimization of the reconstruction error of PCA but this does not necessarily lead to the optimal classification of the face difference images. In the following section we discuss this problem in detail and propose an eigenvalue selection scheme for calculating the residual reconstruction error in PCA or DFFS. Furthermore, we show that the transformational variations among the face images of an individual and noise are always coupled and noise will be magnified when normalized by small eigenvalues in the probabilistic distance measure. To alleviate the problems caused by noise we propose to use PCA to separate the noise from the discriminating features and transformational variations. In this reduced PCA subspace, Bayesian algorithm is applied to remove the effect of transformational variations from the final feature vectors.

3.4 Proposed Method for Eigenvalue Selection in Residual Space

The significant contribution of the Bayesian approach is the estimation of $P(\Delta|\Phi^I)$ which is the sum of the Mahalanobis distance measure and an Euclidean distance equivalent to the PCA residual error [2]. The latter is the difference of the squared original vector length and the squared projected vector length, then weighted by the inverse of the average eigenvalue between m and n as given by (3.4). The complete weight function can be written as

$$w_k = \begin{cases} 1/\lambda_k, & k \leq m \\ 1/\rho, & m < k \leq n \end{cases}, \quad (3.7)$$

where ρ is given by (3.6), k is the index of the eigenvalues. Using the inverse of the eigenvalue to weight the eigenfeature makes a large difference in the face recognition performance. To overcome the problems of zero or small eigenvalues caused by noise and the limited number of training samples a constant average eigenvalue

ρ is used which minimizes the reconstruction error of PCA [87]. However, as m increases, this average eigenvalue tends to become very small value leading to undue high weighting of the reconstruction error in the range from m to n .

For clear illustrations we plot the entire eigenspectrum using the function $\sigma_k = \lambda_k$ and its inverse given by $w_k = 1/\lambda_k$, shown in Fig. 3.2. As this value of k increases the eigenvalues lead to more and more noisy part of the eigenspectrum. For any value $m = k$, the truncation of basis vectors occurs and the entire space is divided into principal subspace (F) and complementary subspace ($\bar{F}$). The DIFS is calculated and weighted according to the weight function w_k and up to the dimensionality of m . Due to noise and limited number of training samples the eigenvalues beyond the range of m becomes very small, if their average is taken the value still becomes low. If its inverse is taken the value becomes very large, weighting the reconstruction error or DFFS with this function results in undue overemphasis in this region, this is illustrated in Fig. 3.2.

Figure 3.2: Weighting functions of (3.6), (3.7) and (3.8) in principal and complementary subspaces based on a typical real eigenspectrum.

This is primarily due to the fact that, in practice, the higher order eigenvalues typically decay very fast and stabilize near the observation noise variance. Actually after the decay of the high valued eigenvalues, noise component dominate the eigenvalue spectrum leading to the faster decay of the variance in the subspace. Therefore, the decay of the eigenvalues should be slowed down to compensate the effect of noise and the finite number of training samples. This can be done by replacing the average eigenvalue with the maximum of the eigenvalues available in the range m and $l - p$, this is given by

$$\rho_{ub} = \max \{ \lambda_i | m \leq i \leq (l - p) \}, \quad (3.8)$$

This suggests a strategy for selecting a constant eigenvalues beyond the range of m which avoids the over-weighting of the reconstruction error or DFFS as shown in Fig. 3.2. We call this ρ_{ub} as ‘upper bound’ ρ since the weighting using the inverse of the average eigenvalue beyond this range tends to be very high leading to the decline of recognition rate. Therefore, the new weighting function can be written as

$$w_k^{ub} = \begin{cases} 1/\lambda_k, & k \leq m \\ 1/\rho_{ub}, & m < k \leq n \end{cases}, \quad (3.9)$$

which is shown in Fig. 3.2.

If we ignore the normalization terms of (3.4), the final distance measure can be written as

$$d(\Delta)_{ML} = \sum_{i=1}^m \frac{y_i^2}{\lambda_i} + \frac{\varepsilon^2(\Delta)}{\rho_{ub}}, \quad (3.10)$$

This ‘upper bound’ selection of ρ attenuates the problems of unreliable small and zero eigenvalues caused by noise and the limited number of training samples. It not only results in improved efficiency and recognition rate but also results in selection of much more stable eigenvalue in the computation of DFFS.

Although the above approach attenuates the noise to some extent but it is still not able to decouple the noise from transformational difference and discriminating features. To solve this problem, we apply PCA to the original image space and by selecting the leading eigenvectors, the noise encoded in the large number of small eigenvalues is removed. This helps us in two ways, firstly, the large dimensionality of the original image is drastically reduced and noises encoded in the image are removed by selecting the eigenvectors corresponding to the largest eigenvalues. Secondly, the intrapersonal variation matrix constructed in the PCA reduced space avoids the large number of trailing small and zero eigenvalues because the rank of the intra-class scatter matrix is always less than the rank of the total-class scatter matrix [4]. Therefore using a two step PCA and Bayesian, we are able to first separate noise from the transformational difference and then transformational difference from the discriminating features. The proposed improved ML approach is then applied in this PCA reduced space.

3.5 Experimental Setup and Results

We tested and compared the proposed ‘upper bound’ ρ (UB_rho) approach on FERET [45] face images with that of average ρ (Av_rho) method. Further, we evaluate the proposed idea of applying PCA first, followed by the Bayesian approach (PCA UB_rho) and also compare with those of the above. The FERET database includes 14,051 face images of 1208 subjects covering a wide range of variations in viewpoints, illuminations, pose, facial expressions and so on [45]. In all the experiments reported in this thesis images are normalized and preprocessed following the CSU Face Identification Evaluation System [73]. In particular, the following preprocessing operations are performed: (1) Images are rotated and scaled so that the centers of the eyes are placed on the specific pixels (same horizontal axis). The normalization crops the image to $h \times w$ pixels with fixed number of pixels between the centers of the eyes. (2) A standard mask is applied to remove

non-face portions. (3) Histogram equalization is performed in the no-masked portion of the face and image intensity values are normalized to zero mean and unit standard deviation. (4) Using lexicographic ordering, each image of size $h \times w$ is finally represented as a column vector of dimensionality $n \times 1$, where $n = h \times w$. Fig. 3.3 shows sample images from the FERET database for four subjects with two sample images per subject and their normalized versions.

Figure 3.3: Sample images from the FERET database for four subjects with two sample images per subject, mask and their normalized versions.

In all experiments reported in this section, the images are cropped into the size

of 38×33 . FERET dataset consists of 2388 frontal images comprising of 1194 persons (two images FA/FB per person) is selected. We divide this database into gallery and testing datasets, for all the experiments reported here, there is no overlap in subjects between the training and testing sets.

In the first experiment, we use the recognition rate versus feature number used in the matching, this kind of curve is commonly used by many recent researchers [44, 89, 88] for testing and comparing their algorithms. From the above database, images of 497 subjects are randomly chosen for training and remaining 697 subjects are used for testing. Therefore, 697 images in the gallery and 697 images in the probe dataset. Recognition rate is the ratio between the number of first correct matches between the probe images with the gallery images and the total number of probe images. In all our experiments, while applying PCA, the dimensionality of the original images are reduced to 300 feature vectors. We also tested traditional PCA and FDA techniques in our experiments, since their performances are far below that of the Bayesian approach we do not present them in our results.

Fig. 3.4 shows the performance of our proposed approaches as compared to the existing one. From the figure, it is clearly visible that the performance of the existing Av_rho Bayesian methodology decreases very rapidly with the increasing number of feature vectors. This is because of its inability of separating the noise encoded in the eigenvectors corresponding to the smaller eigenvalues and also its lack of ability for generalization over the tested dataset. Whereas, our proposed methodologies of UB_rho and PCA followed by UB_rho show more stability and better generalization ability over the used feature numbers. We also plot PCA Av_rho methodology, whose performance is lower than that of our proposed PCA UB_rho Bayesian approach. From Fig. 3.4 we see that the proposed approaches consistently outperform the existing approach (Av_rho) for all number of features.

In the second part of the experiment, we perform extensive evaluation of our proposed methods using recognition rate versus recognition rank plots for various number of feature numbers and decreasing number of samples in the training datasets.

Figure 3.4: Recognition rate against the number of features used in the matching on the FERET database of 994 training images (497 subjects) and 1394 testing images (697 subjects).

This analysis tool is called AnalyzeRankCurve and is used in [73, 45, 2, 88], it helps in evaluating the overall performance and efficiency of different algorithms. For this part, we conduct experiments on decreasing number of training image samples. Firstly, images of 700 subjects (with two images per subject) are randomly chosen for training and remaining 494 subjects are used for testing. Hence, the gallery and probe datasets include 494 images each. Fig. 3.5 shows the recognition rate on the testing set against the rank calculated in the matching process.

Recognition rank can be interpreted as follows: A one means that for the corresponding test image and distance measure, the first match of a gallery image of the same subject is found at position one in the gallery sorted by increasing distance relative to the test image. This means that one means the best match is of the same subject. Likewise, a two means there is a one gallery image of another subject that is closer than the gallery image of the same subject and so on. This value is commonly called the recognition rank. These graphs are plotted for recognition rate against rank for different selection of feature numbers. Although in these graphs, the recognition performance for all the ranks are important for overall evaluation of the algorithm, the recognition performance of rank one is our prime interests, since it is the top one match on the testing set. It is evident that our approaches perform significantly better than the existing method for this rank one match. Also, from Fig. 3.5 we see that the proposed UB_rho and PCA UB_rho approaches consistently outperform the existing method for all number of features and ranks.

In the next experiment we select fewer training samples. Images of 497 subjects are randomly chosen for training and remaining 697 subjects are used for testing. Therefore, 697 images in the gallery and 697 images in the probe dataset. Fig. 3.6 shows the performance of our proposed methods for decreasing number of feature numbers used for matching. It demonstrates again that the proposed UB_rho and PCA UB_rho approaches consistently outperform the existing method for all number of features and ranks.

Figure 3.5: Recognition rate against the rank plots for various number of feature vectors (a) $m=200$, (b) $m=150$, (c) $m=100$, (d) $m=50$ used in the matching of FERET database of 1400 training images (700 subjects) and 988 testing images (494 subjects).

Figure 3.6: Recognition rate against the rank plots for various number of feature vectors (a) $m=200$, (b) $m=150$, (c) $m=100$, (d) $m=50$ used in the matching of FERET database of 994 training images (497 subjects) and 1394 testing images (697 subjects).

To evaluate the recognition performance on further smaller number of training samples, images of 250 subjects are randomly selected for training and the remaining images of 944 subjects are used for testing. Fig. 3.7 shows the recognition rate on the testing set against the rank used in the matching process. It demonstrates

Figure 3.7: Recognition rate against the rank plots for various number of feature vectors (a) $m=200$, (b) $m=150$, (c) $m=100$, (d) $m=50$ used in the matching of FERET database of 500 training images (250 subjects) and 1888 testing images (944 subjects).

once more that the proposed UB_rho and PCA UB_rho approaches consistently outperform the existing method for all number of features and ranks.

It is more plausible that different training datasets will give different recognition performances over the same gallery and probe datasets. Also, the selection of the number of feature vectors used in the matching process can make significance

difference in the overall recognition performances. Hence, for a clearer comparison of these approaches, we extract the average recognition rate of the top 50% of total number of features used obtained from the first part of the experiment. This will help us to understand that, in general, up to what extent our approaches can outperform the existing method. In our case total number of features we tested is 240. Table 3.1 shows the recognition rate on the testing set against the methodologies used in the experiments. It clearly shows the superior performance

Table 3.1: Recognition Rate (%) of Different Approaches on Different Data Sizes

method	sizes of training/testing set		
	1400 / 988	994 / 1394	500 / 1888
Av_rho	89.70	90.44	89.76
UB_rho	93.64	93.03	91.03
PCA UB_rho	96.46	95.94	92.90

of the proposed UB_rho and PCA UB_rho approaches to the existing Bayesian approach.

3.6 Summary

This section addresses the problems of weighting the eigenfeatures of face images based on a set of training samples. Using the inverse of the eigenvalue to weight the eigenfeature makes large difference in the face recognition performance. To overcome the problems of weighting by the inverse of small or zero eigenvalues we propose an ‘upper bound’ ρ selection mechanism, which is applied to improve the performance of Bayesian approach for face recognition. This ‘upper bound’ selection of ρ attenuates the problems of unreliable small and zero eigenvalues caused by noise and the limited number of training samples. Especially for small or zero eigenvalues, their inverses are highly sensitive to noise disturbance and the effect of limited number of training samples. To overcome these difficulties, we also propose to apply Bayesian approach in the reduced PCA subspace. Using this approach, we are able to first separate noise from the transformational difference

and then transformational difference from the discriminating features. The new approach is much less sensitive to noises and achieves better generalization or less over-fitting the training data than the standard Bayesian approach. Experimental results of proposed approaches on various number of training samples on FERET dataset demonstrate the improvement over the existing method.

Chapter 4

Eigenfeature Regularization and Extraction for Face Recognition in Linear Space

4.1 Introduction

Face recognition has attracted many researchers in the area of pattern recognition, computer vision and machine learning because of its immense application potential. Over the last two decades numerous methods have been proposed and have demonstrated good performances in terms of accuracy and robustness [3, 7]. However, there are still substantial challenging problems remain to be solved. One of the critical issues is how to extract discriminative and stable features for classification. Recently, three statistical techniques for subspace feature extraction have been extensively studied and have become very popular in face recognition. These are principal component analysis (PCA) [26], Bayesian maximum likelihood (ML) [44, 2, 87] and Fisher discriminant analysis (FDA) [31, 16]. A theoretical analysis showed that a low-dimensional linear subspace could capture the set of images of an object produced by a variety of lighting conditions [90]. The agreeable proper-

ties of the linear subspace analysis and its promising performance achieved in the face recognition encourage researchers extending it to the higher order statistics [91, 92], nonlinear methods [93, 94, 95], Gabor features [96, 97] and the locality preserving projections [15, 98]. However, the basic linear subspace analysis has still outstanding challenging problems when applied to the face recognition due to the high dimensionality of the face image and the finite number of training samples in practice.

PCA uses covariance matrices of the training data to find a linear subspace efficient for representation. Using the covariance matrix it finds a subspace that minimizes the mean-square error (MSE) between the feature vectors and their projections in this space. Based on the covariance matrix, the data mean is effectively shifted to the origin of the feature space. The MSE solution lines up with the primary directions of the data which de-correlates the feature data via rotation. This is optimal for representation or reconstruction of data point of view. However, the best representation of data might not be useful for classification. To differentiate face images of one person from those of the others, the discrimination of the features is the most important. FDA is an efficient way to extract the discriminative features as it handles the within- and between-class variations separately. However, this method needs the inverse of the within-class scatter matrix. This is problematic in many practical face recognition tasks because the dimensionality of the face image is usually very high compared to the number of available training samples and hence the within-class scatter matrix is often singular.

Numerous methods have been proposed to solve this problem in the last decade. A popular approach called Fisherface [32] applies PCA first for dimensionality reduction so as to make the within-class scatter matrix nonsingular before the application of FDA. However, applying PCA for dimensionality reduction may lose discriminative information [35, 88, 14]. Direct-LDA (DLDA) method [34, 89] removes null space of the between-class scatter matrix and extracts the eigenvectors corresponding to the smallest eigenvalues of the within-class scatter matrix. It is

an open question of how to scale the extracted features as the smallest eigenvalues are very sensitive to noise. To overcome the problems of small and zero eigenvalues, Lu *et al.* proposed direct fractional-step LDA method, which uses modified Fisher criteria to obtain low dimensional discriminative features [89]. However, Gao *et al.* have shown that direct-LDA is not equivalent to LDA [99]. The null space (NLDA) approaches [12, 38, 14] assume that the null space contains the most discriminative information. Interestingly, this appears to be contradicting the popular FDA that only uses the principal space and discards the null space. A common problem of all these approaches is that they all lose some discriminative information, either in the principal or in the null space.

In practice, both the subspaces have distinct roles to play in the overall accuracy of a face recognition system. To use the both subspaces, a modified FDA approach [100] replaces the within-class scatter matrix by the total scatter matrix. The following work [101] extracts features separately from the principal and null spaces of the within-class scatter matrix. However, the extracted features may not be properly scaled and undue over-emphasis is placed on the null space in these two approaches due to the replacement of the within-class scatter matrix by the total scatter matrix. The dual-space approach (DSL) [88] scales features in the complementary subspace by the average eigenvalue of the within-class scatter matrix over this subspace. As eigenvalues in this subspace are not well estimated [88], their average may not be a good scaling factor relative to those in the principal subspace. Features extracted from the two complementary subspaces are properly fused by using summed normalized-distance [95]. There are two problems associated with these approaches: how to divide the space into the principal and the complementary subspaces and how to apportion a given number of features to the two subspaces. Moreover, these subspaces have distinct roles in the overall recognition rates, thus it is inefficient or only suboptimal to extract features separately from the two subspaces.

The above approaches focus on the problem of singularity of the within-class

scatter matrix. In fact, the instability and noise disturbance of the small eigenvalues cause great problems when the inverse of matrix is applied such as in the Mahalanobis distance, in the ML estimation and in the whitening process of various FDA approaches. Problems of the noise disturbance were addressed in [13] and a unified framework of subspace methods (UFS) was proposed. The good recognition performance of this framework shown in [13] verifies the importance of the noise suppression. However, this approach applies three stages of subspace decompositions sequentially on the face training data and the dimensionality reduction occurs at the very first stage. As addressed in the literature [35, 88, 14], applying PCA for dimensionality reduction may lose discriminative information. Another open question of UFS is how to choose the number of principal dimensions for the first two stages of subspace decompositions before selecting the final number of features in the third stage. The experimental results in [13] show that the recognition performance is sensitive to these choices at different stages.

In this chapter, we present a new approach for facial eigenfeature regularization and extraction. Image space spanned by the eigenvectors of the within-class scatter matrix is decomposed into three subspaces. Eigenfeatures are regularized differently in these subspaces based on an eigenspectrum model. This alleviates the problems of unreliable small and zero eigenvalues caused by noise and the limited number of training samples. It also enables the discriminant evaluation performed in the full dimension of the image data. Feature extraction or dimensionality reduction occurs only at the final stage after the discriminant assessment. In the next section, we first model the eigenspectrum, study the effect of the unreliable small eigenvalues on the feature weighting and decompose the eigenspace into face-, noise- and null-subspaces. Eigenfeature regularization and extraction are presented in section 4.3. Analysis of the proposed approach and comparison with other relevant methods are provided in section 4.4. Experimental results are presented in section 4.5. Summary is drawn in section 4.6.

4.2 Eigenspectrum Modeling and Subspace Decomposition

From the definitions stated in chapter 3, the within-class scatter matrix is defined by

$$\mathbf{S}^w = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T, \quad (4.1)$$

where $\bar{X}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} X_{ij}$. Let the training set contain p persons and q_i sample images for person i . The number of total training sample is $l = \sum_{i=1}^p q_i$. ‘ T ’ as superscript denotes transpose of the matrix. The between-class scatter matrix $\mathbf{S}^b$ and the total (mixture) scatter matrix $\mathbf{S}^t$ are defined by

$$\mathbf{S}^b = \sum_{i=1}^p c_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T, \quad (4.2)$$

$$\mathbf{S}^t = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})^T, \quad (4.3)$$

where $\bar{X} = \sum_{i=1}^p c_i \bar{X}_i$. If all classes have equal prior probability, then $c_i = 1/p$, such that $\sum_{i=1}^p c_i = 1$.

Let $\mathbf{S}^g, g \in \{t, w, b\}$ represent one of the above scatter matrices. If we regard the elements of the image vector or the class mean vector as features, these preliminary features will be de-correlated by solving the eigenvalue problem

$$\mathbf{\Lambda}^g = \mathbf{\Phi}^{gT} \mathbf{S}^g \mathbf{\Phi}^g, \quad (4.4)$$

where $\mathbf{\Phi}^g = [\phi_1^g, \dots, \phi_n^g]$ is the eigenvector matrix of $\mathbf{S}^g$, and $\mathbf{\Lambda}^g$ is the diagonal matrix of eigenvalues $\lambda_1^g, \dots, \lambda_n^g$ corresponding to the eigenvectors.

Suppose that the eigenvalues are sorted in descending order $\lambda_1^g \geq \dots \geq \lambda_n^g$. The plot of eigenvalues λ_k^g against the index k is called eigenspectrum of the face

training data. It plays a critical role in the subspace methods as the eigenvalues are used to scale and extract features. We first model the eigenspectrum to show its problems in the feature scaling and extraction.

4.2.1 Eigenspectrum Modeling

If we regard X_{ij} as samples of a random variable vector X , the eigenvalue λ_k^g is a variance estimate of X projected on the eigenvector ϕ_k^g estimated on the training samples. It usually deviates from the true variance of the projected random vector X due to the finite number of training samples. Thus, we model the eigenspectrum in the range subspace $\lambda_k^g, 1 \leq k \leq r$, as the sum of the true variance component v_k^F and a deviation component δ_k . For simplicity, we call v_k^F face component and δ_k noise component. As the face component typically decays rapidly and stabilizes, we can model it by a function of the form $1/f$ that can well fit to the decaying nature of the eigenspectrum. The function form $1/f$ was used in [44] to fit the eigenspectrum in the whole range subspace and then to extrapolate eigenvalues in the null space for computing the average eigenvalue over a subspace that includes the null space. The noise component δ_k that includes the effect of the finite number of training samples can be negative if the face component v_k^F is modeled by a $1/f$ function that always has positive values.

We propose to model the eigenvalues first in descending order of the face component v_k^F by

$$\hat{\lambda}_k^F = v_k^F + \delta_k = \frac{\alpha}{k + \beta} + \delta_k, \quad 1 \leq k \leq r \quad (4.5)$$

where α and β are two constants that will be given in the next section. The modeled eigenspectrum $\hat{\lambda}_k^g$ is then obtained by sorting $\hat{\lambda}_k^F$ in descending order. As the eigenspectrum decays very fast, we plot the square roots $\sigma_k^g = \sqrt{\lambda_k^g}$ and $\hat{\sigma}_k^g = \sqrt{\hat{\lambda}_k^g}$ for clearer illustration (we still call them eigenspectrum for simplicity). A typical real eigenspectrum σ_k^g and its model $\hat{\sigma}_k^g$ are shown in Fig. 4.1, where δ_k

is a random number evenly distributed in a small range of $[-\lambda_1^g/1000, \lambda_1^g/1000]$. From Fig. 4.1 we see that the proposed model closely approximates to the real eigenspectrum.

4.2.2 Problems of Feature Scaling and Extraction

Our findings and also in [102, 89] suggest that in face recognition, the application of PCA step does not guarantee the successful application of FDA as the subsequent step. The transformed within-class scatter matrix might still be singular. This occurs because of the effect of small/finite number of training samples, high dimensionality of face images and noises. The PCA approach with Euclidean distance selects d leading eigenvectors and discards the others. This can be seen as weighting the eigenfeatures with a step function as

Figure 4.1: A typical real eigenspectrum, its model and feature weighting/extraction of PCA (left); the first 50 real eigenvalues and the values of the model (right).

$$u_k^d = \begin{cases} 1, & k \leq d \\ 0, & k > d \end{cases}. \quad (4.6)$$

This weighting function of PCA $w_k^P = u_k^d$ is shown in Fig. 4.1. In a practical face recognition problem, the recognition performance usually improves with the increase of d . The PCA with Mahalanobis distance can be seen as the PCA with

Euclidean distance and a weighting function $w_k^M = u_k^d / \sigma_k^g$. This weighting function is shown in Fig. 4.1. Using the inverse of the square root of the eigenvalue to weight the eigenfeature makes a large difference in the face recognition performance. The recognition accuracy usually increases sharply with the increase of d and is better than the Euclidean distance for smaller d . But it decreases also sharply and is much worse than the Euclidean distance for larger d . The ML approach that weights the principal features ($1 \leq k \leq d$) by the inverse of the square root of the eigenvalue also suffers from the performance decrease with the increase of the d after d reaches a small value ($d \ll r$).

Noise disturbance and poor estimates of small eigenvalues due to the finite number of training samples are the culprits. The limited number of training samples results in very small eigenvalues in some dimensions that may not well represent the true variance in these dimensions. This may result in serious problems if their inverses are used to weight the eigenfeatures. The characteristics of the eigenspectrum and the generalized deterioration caused by the small eigenvalues were well addressed in [103]. To exclude the small eigenvalues in the discriminant evaluation, probabilistic reasoning models and enhanced FDA models were proposed and compared in [103]. The modeled eigenspectrum in Fig. 4.1 that approximates closely to the real one is re-sorted in descending order of the face component v_k^F and is plotted in Fig. 4.2. The small noise disturbances are now visible in Fig. 4.2 given that the variances of the face component should always decay. These small disturbances cause large vibrations of the inverse eigenspectrum as shown in Fig. 4.2. Eigenfeatures of larger index k are heavily weighted by the scaling factors that are highly sensitive to noise and training data. This causes the deterioration of the recognition performance, especially on the independent testing data.

4.2.3 Subspace Decomposition

Subspace decomposition is an important issue in face recognition problems because as compared to other computer vision problems face images are of very high

Figure 4.2: A typical real eigenspectrum, its model sorted in descending order of the face component and their inverse; decomposition of the eigenspace into face-, noise- and null-subspaces.

dimensions and are often very difficult to deal with. Since appearance/subspace based face recognition use whole face image the covariances matrices are quite large. How to handle this large eigenspace produced by these covariance matrices, so that effective discriminant features can be extracted is an open question in the existing literature. Presently, most of the algorithms decompose the whole eigenspace of various scatter matrices using their ranks [12, 104]. Some authors also decomposed by trial and error or heuristics, examples include [95, 13, 88].

From our previous discussion, as shown in Fig. 4.2, that the small noise disturbances have little effect on the initial portion of the eigenspectrum cause large vibrations of the inverse eigenspectrum in the region of small eigenvalues. Therefore, we propose to decompose the eigenspace $\mathbb{R}^n$ spanned by eigenvectors $\{\phi_k^g\}_{k=1}^n$ into three subspaces: a reliable face variation dominating subspace (or simply face space) $\mathbf{F} = \{\phi_k^g\}_{k=1}^m$, an unstable noise variation dominating subspace (or simply noise space) $\mathbf{N} = \{\phi_k^g\}_{k=m+1}^r$ and a null space $\emptyset = \{\phi_k^g\}_{k=r+1}^n$ as illustrated in Fig. 4.2. The purpose of this decomposition is to modify or regularize the unreliable eigenvalues for better generalization.

It is not difficult to determine the rank of the scatter matrix r . For $\mathbf{S}^t$, $r \leq \min(n, l - 1)$; for $\mathbf{S}^w$, $r \leq \min(n, l - p)$; and for $\mathbf{S}^b$, $r \leq \min(n, p - 1)$. In practice, the rank of a scatter matrix usually reaches the corresponding one of these maximum values unless some training images are linearly dependent. Even in this rare case, the rank r can be easily determined by finding the maximal value of k that satisfies $\lambda_k^g > \varepsilon$ where ε is a very small positive value comparing to λ_1^g .

As face images have similar structure, significant face components reside intrinsically in a very low-dimensional (m -dimensional) subspace. For a robust training, the database size should be significantly larger than the face dimensionality m , although it could be, and usually in practice is, much smaller than the image dimensionality n . Thus, in many practical face recognition training tasks we usually have $m \ll r \ll n$. As the face component typically decays rapidly and stabilizes, eigenvalues in the face dominant subspace, which constitute the initial portion of the eigenspectrum, are the outliers of the whole spectrum. It is well known that median operation works well in separating outliers from a data set. To determine the start point of the noise dominant region $m + 1$, we first find a point near the center of the noise region by

$$\lambda_{med}^g = \text{median} \{ \forall \lambda_k^g \mid k \leq r \}. \quad (4.7)$$

The distance between λ_{med}^g and the smallest nonzero eigenvalue is $d_{m,r} = \lambda_{med}^g - \lambda_r^g$. The upper bound of the unreliable eigenvalues is estimated by $\lambda_{med}^g + d_{m,r}$. Although this is a reasonable choice of the upper bound of the unreliable eigenvalues, it may not be optimal in all cases considering the great variation of image size and number of training samples in different applications. More generally, the start point of the noise region $m + 1$ is estimated by

$$\lambda_{m+1}^g = \max \{ \forall \lambda_k^g \mid \lambda_k^g < (\lambda_{med}^g + \mu(\lambda_{med}^g - \lambda_r^g)) \}. \quad (4.8)$$

where μ is a constant. The optimal value of μ may be slightly larger or smaller

than 1 for different applications. To avoid exhaustive search for the best parameter value, μ is fixed to be 1 in all experiments of this chapter for fair comparisons with other approaches.

4.3 Eigenfeature Regularization and Extraction

Many researchers have been dedicated to searching for more effective discriminant subspaces [34, 101, 100, 38, 105, 106]. Essentially this means finding or extracting features from eigenspace which are effective for discrimination purposes. In general, it is desired to extract features that have the smallest within-class variations and the largest between-class variations. Problems occur in seeking the smallest within-class variations because the variances are estimated based on the finite number of training samples and hence the estimated smaller variances are unstable and tend to over-fit the specific training data. It is not a surprise that the within-class variation is the biggest obstacle to achieve high recognition rate. Thus, we first work on the within-class scatter matrix $\mathbf{S}^w$.

After solving the eigenvalue problem as (4.4), a unit within-class scatter matrix in the subspace $\bar{Y}_{ij} \in \mathbb{R}^r$ can be obtained by representing training samples with new feature vectors $\bar{Y}_{ij}$

$$\bar{Y}_{ij} = \bar{\Phi}_r^{wT} X_{ij}, \quad (4.9)$$

where $\bar{\Phi}_r^w = [\phi_1^w/\sigma_1^w, \dots, \phi_r^w/\sigma_r^w]$ are composed of so called whitened eigenvectors of $\mathbf{S}^w$ with nonzero eigenvalues and $\|\phi_k^w\| = 1$. Thus, an n -dimensional image vector X_{ij} is first represented by an n -dimensional eigenfeature vector $Y_{ij} = \Phi^{wT} X_{ij}$, and then multiplied by a weighting function

$$w_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq r \\ 0, & r < k \leq n \end{cases}, \quad (4.10)$$

as shown in Fig. 4.3. Training data represented by the new feature vectors $\bar{Y}_{ij}$ will produce the same (unit) within-class variances in all directions of the reduced feature space $\mathbb{R}^r$. It appears that the problem of the within-class variation is solved.

However, there are two problems. Firstly, face structural information in the null space is lost, or equivalently, features in the null space is weighted by a constant zero. This is unreasonable because features in the null space have zero within-class variances based on the training data and hence should be much heavier weighted. It seems anomalous that the weighting function increases with the decrease of the eigenvalues and then suddenly has a big drop from the maximum value to zero as shown in Fig. 4.3. Secondly, weights determined by the inverse of σ_k^w is, though optimal in terms of the ML estimation, dangerous when σ_k^w is small. This is shown in the modeled eigenspectrum in Fig. 4.2. The small and zero eigenvalues are training-set-specific and very sensitive to different training sets. Adding new samples to the training set or using different training set may easily change some zero eigenvalues to nonzero and make some very small eigenvalues several times larger. Therefore, eigenspectrum needs to be regularized.

4.3.1 Eigenspectrum Regularization

Although there is always noise component in $\mathbf{F}$ as noise affects every element of the image vector, its variance is very small comparing to the large variance of the face structural component in $\mathbf{F}$. In $\mathbf{N}$, however, noise component may dominate in the variance changes and the finite number of training samples results in faster decay of the variances. Therefore, the decay of the eigenvalues should be slowed down to compensate the effect of noise and the finite number of training samples. This can be done by replacing the eigenspectrum with the proposed model (4.5).

As the eigenspectrum in the face space is dominated by the face structural component, the parameters of α and β are determined by fitting the face portion

of the model $v_k^F = \alpha/(k + \beta)$ to the real eigenspectrum in the face space $\mathbf{F}$. While not limiting ourselves from other possible fitting methods, in all experiments of this work, we simply determine α and β by letting $v_1^F = \lambda_1^w$ and $v_m^F = \lambda_m^w$, which yields

$$\alpha = \frac{\lambda_1^w \lambda_m^w (m - 1)}{\lambda_1^w - \lambda_m^w}, \quad (4.11)$$

$$\beta = \frac{m \lambda_m^w - \lambda_1^w}{\lambda_1^w - \lambda_m^w}. \quad (4.12)$$

Fig. 4.3 shows the square roots of a real eigenspectrum σ_k^w and the face portion of its model $\sigma_k^F = \sqrt{v_k^F}$. We see that the model σ_k^F fits closely to the real σ_k^w in the face space $\mathbf{F}$ but has slower decay in the noise space $\mathbf{N}$. The faster decay of the real eigenspectrum σ_k^w in $\mathbf{N}$ due to noise and the effect of the limited number of training samples is what we want to slow down.

Figure 4.3: Weighting functions of (4.10) and (4.14) in the face-, noise- and null-subspaces based on a typical real eigenspectrum.

In the null space, we have no information about the variation of the eigenvalues and hence all features are treated in a same way. The zeros variance in the null space is only an estimate on one set of the training data. Another set of training data may easily make them nonzero, especially when larger number of training

samples are used. Therefore, we should not trust the zero variance and derive an infinite or very large feature weights in this space. However, based on the available training data that result in zero variances in the null space, the feature weights in the null space should not be smaller than those in the other subspaces.

Therefore, we regularize the eigenspectrum by replacing the noise dominating λ_k^w in $\mathbf{N}$ with the face portion of the model $v_k^F = \alpha/(k + \beta)$ and replacing the zero λ_k^w in null space $\mathcal{O}$ with the constant v_{r+1}^F . Thus, the regularized eigenspectrum $\tilde{\lambda}_k^w$ is given by

$$\tilde{\lambda}_k^w = \begin{cases} \lambda_k^w, & k < m \\ \frac{\alpha}{k+\beta}, & m \leq k \leq r \\ \frac{\alpha}{r+1+\beta}, & r < k \leq n \end{cases} \quad (4.13)$$

The proposed feature weighting function is then

$$\tilde{w}_k^w = \frac{1}{\sqrt{\tilde{\lambda}_k^w}}, \quad k = 1, 2, \dots, n. \quad (4.14)$$

Fig. 4.3 shows the proposed feature weighting function $\tilde{w}_k^w$ calculated by (4.11), (4.12), (4.13) and (4.14) comparing with that w_k^w of (4.10). Obviously, the new weighting function $\tilde{w}_k^w$ is identical to w_k^w in the face space, increases along with k at a much slower pace than w_k^w in the noise space and has maximal weights instead of zero of w_k^w in the null space.

Using this weighting function and the eigenvectors ϕ_k^w , training data are transformed to

$$\tilde{Y}_{ij} = \tilde{\Phi}_n^{wT} X_{ij}, \quad (4.15)$$

where

$$\tilde{\Phi}_n^w = [\tilde{w}_k^w \phi_k^w]_{k=1}^n = [\tilde{w}_1^w \phi_1^w, \dots, \tilde{w}_n^w \phi_n^w] \quad (4.16)$$

is a full rank matrix that transforms an image vector to an intermediate feature vector. There is no dimension reduction in this transformation as $\tilde{Y}_{ij}$ and X_{ij} have the same dimensionality n .

4.3.2 Eigenfeature Extraction

After performing the feature regularization in the previous section, a new total scatter matrix is formed by vectors $\tilde{Y}_{ij}$ of the training data as

$$\tilde{\mathbf{S}}^t = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} (\tilde{Y}_{ij} - \bar{Y})(\tilde{Y}_{ij} - \bar{Y})^T, \quad (4.17)$$

where $\bar{Y} = \sum_{i=1}^p \frac{c_i}{q_i} \sum_{j=1}^{q_i} \tilde{Y}_{ij}$.

There is some difference between the total scatter matrix $\tilde{\mathbf{S}}^t$ and the between-class scatter matrix $\tilde{\mathbf{S}}^b$ of the training data $\tilde{Y}_{ij}$ because the within-class scatter matrix is not fully whitened in the noise space $\mathbf{N}$ and the null space $\emptyset$ is not discarded. Some works [33, 98] show that, when the training data is small, PCA can outperform LDA and that PCA is less sensitive to the different training databases. Our experiments show that $\tilde{\mathbf{S}}^t$ outperforms $\tilde{\mathbf{S}}^b$, but only very marginally and not consistently for different numbers of features. For a training set that contains images of only few people, $\tilde{\mathbf{S}}^b$ may have problem of extracting sufficient number of features. The maximal number of features extracted based on $\tilde{\mathbf{S}}^b$ is $p - 1$ at most while that based on $\tilde{\mathbf{S}}^t$ could be much larger ($\sum_{i=1}^p q_i - 1$). Thus, in this work, we suggest to employ the total scatter matrix $\tilde{\mathbf{S}}^t$ of the regularized training data to extract the discriminative features.

The regularized features $\tilde{Y}_{ij}$ will be de-correlated for $\tilde{\mathbf{S}}^t$ by solving the eigenvalue problem as (4.4). Suppose that the eigenvectors in the eigenvector matrix $\tilde{\Phi}_n^t = [\tilde{\phi}_1^t, \dots, \tilde{\phi}_n^t]$ are sorted in descending order of the corresponding eigenvalues. The dimensionality reduction is performed here by keeping the eigenvectors with the d

largest eigenvalues

$$\tilde{\Phi}_d^t = [\tilde{\phi}_k^t]_{k=1}^d = [\tilde{\phi}_1^t, \dots, \tilde{\phi}_d^t], \quad (4.18)$$

where d is the number of features usually selected by a specific application.

Thus, the proposed feature regularization and extraction matrix $\mathbf{U}$ is given by

$$\mathbf{U} = \tilde{\Phi}_n^w \tilde{\Phi}_d^t, \quad (4.19)$$

which transforms a face image vector X , $X \in \mathbb{R}^n$, into a feature vector F , $F \in \mathbb{R}^d$, by

$$F = \mathbf{U}^T X. \quad (4.20)$$

We see that the decomposition of the image space into three subspaces is used only for the regularization of the feature scaling. The discriminant evaluation (here the evaluation of the eigenvalues of $\tilde{\mathbf{S}}^t$) is performed in the full space $\mathbb{R}^n$. Thus, feature extraction is not restricted to project an image vector into one or two of these three subspaces. More specifically, any single feature in F is extracted from the whole space $\mathbb{R}^n$ since any final projection vector in $\mathbf{U}$ may have nonzero components in all the three subspaces.

4.3.3 The Proposed Algorithm

The proposed eigenfeature regularization and extraction (ERE) approach is summarized below:

At the training stage:

- (1) Given a training set of face image vectors $\{X_{ij}\}$, compute $\mathbf{S}^w$ using (4.1) and solve the eigenvalue problem as (4.4).

- (2) Decompose the eigenspace into face-, noise-, and null-space by determining the m value using (4.7) and (4.8).
- (3) Transform the training samples represented by X_{ij} into $\tilde{Y}_{ij}$ using (4.15) with the weighting function (4.14) determined by (4.11), (4.12) and (4.13).
- (4) Compute $\tilde{\mathbf{S}}^t$ using (4.17) with $\tilde{Y}_{ij}$ and solve the eigenvalue problem as (4.4).
- (5) Obtain the final feature regularization and extraction matrix using (4.16), (4.18) and (4.19) with a predefined number of features d . #

At the recognition stage:

- (1) Transform each n -D face image vector X into d -D feature vector F using (4.20) and the feature regularization and extraction matrix $\mathbf{U}$ obtained in the training stage.
- (2) Apply a classifier trained on the gallery set to recognize the probe feature vectors. #

In the experiments of this work, a simple first nearest neighborhood classifier (1-NNK) is applied to test the proposed eigenfeature regularization and extraction (ERE) approach. Cosine distance measure between a probe feature vector F_P and a gallery feature vector F_G

$$dst(F_P, F_G) = -\frac{F_P^T F_G}{\|F_P\|_2 \|F_G\|_2} \quad (4.21)$$

is applied to the proposed approach, where $\|\cdot\|_2$ is the norm 2 operator.

4.4 Analysis and Comparison

In this section we analyze our proposed method with the current state-of-art methodologies. Two techniques are developed and integrated in the proposed

ERE approach for face recognition. Firstly, it regularizes the real eigenspectrum using the novel modeled eigenvalues and performs dimensionality reduction or feature extraction only at the final stage after the discriminant assessment. Secondly, the proposed eigenspectrum regularization not only facilitates this discriminative information search in the whole space but also alleviates the over-fitting problem.

4.4.1 Discriminant Evaluation in a Subspace, Complementary Subspaces and the Whole Space

Most subspace approaches such as FDA, DLDA, UFS and various null space approaches discard a subspace before the discriminant evaluation. The extracted features are only suboptimal as they are the most discriminative only in a subspace.

Although ML works in the whole space, it does not evaluate the discriminant value and hence the whole face image must be used in matching. A modified LDA approach [100] estimates the discriminant value in the full space of $\mathbf{S}^t$ by replacing $\mathbf{S}^w$ with $\mathbf{S}^t$. It is easy to see that, after whitening by $\mathbf{S}^t$, the between-class variance is always one in all dimensions of the null space of $\mathbf{S}^w$ but is always smaller than one in all dimensions of the range space of $\mathbf{S}^w$. Obviously, the null space of $\mathbf{S}^w$ is unduly over-emphasized and the extracted features may not be properly scaled due to the replacement of $\mathbf{S}^w$ by $\mathbf{S}^t$ in the discriminant assessment.

Some approaches [101, 88, 95] extract features separately from the two complementary subspaces of $\mathbf{S}^w$. One of them [101] first extracts all the $p - 1$ features from the null space of $\mathbf{S}^w$. It then extracts some features from the range space of $\mathbf{S}^w$ where $\mathbf{S}^t$ is used as a replacement of $\mathbf{S}^w$ in the discriminant evaluation. Obviously, this method, similar to the approach in [100], over-emphasizes the null space and may not properly scale the features. The DSL approach [88] scales features in the complementary subspace by the average eigenvalue of $\mathbf{S}^w$ over this subspace. As the eigenvalues in this subspace are not well estimated [88], their

average may not be a proper scaling factor relative to those in the principal subspace. To circumvent the feature scaling problem in the training phase, a summed normalized-distance is proposed in [95] that properly combines the two groups of features in the recognition phase.

Although these double-space approaches do not throw away any subspace before the discriminant evaluation, it is inefficient or only suboptimal to evaluate the discriminant value and extract features separately in two subspaces. Other open questions include how to divide the space into the two subspaces properly and how to apportion the given number of features to the two subspaces reasonably. In contrast, the proposed ERE approach searches the most discriminative information in the whole space and the feature scaling problem caused by the small and zero eigenvalues is alleviated by the eigenspectrum regularization.

4.4.2 Eigenspectrum Regularization

It is well-known that the estimated covariance matrix or its eigenvalues need to be regularized, especially for small number of training samples. Problems of the biased estimation of eigenvalues that cause over-fitting and hence poor generalization were well addressed in [107].

A method called RD-QDA [108] was proposed to regularize $\mathbf{S}^w$ and applied in the face recognition. Another regularization approach called R-LDA [109] was derived by replacing $\mathbf{S}^w$ with $\mathbf{S}^w + \eta \mathbf{S}^b$, where η is a regularization parameter. However, the both approaches are based on the DLDA framework and the appropriate values of the regularization parameter are different for different databases and database sizes. These critical values are determined by exhaustive search with the help of a validation data set in the experiment. Furthermore, the both approaches regularize the pooled within-class scatter matrix equivalently by adding a constant to all eigenvalues. Although the largest sample-based eigenvalues are biased high and the smallest ones are biased low, as pointed out in [107], the bias is most pronounced

when the population eigenvalues tend toward equality, and it is correspondingly less severe when their values are highly disparate. For the application of face recognition, it is well-known that the eigenspectrum first decays very rapidly and then stabilizes. Hence, adding a constant to the eigenspectrum may bias back the rapidly changing eigenvalues in $\mathbf{F}$ too much that introduces additional error source, and bias back the flat eigenvalues in $\mathbf{N}$ too little at the same time.

Therefore, the proposed approach decomposes the image space into a highly disparate face dominant subspace $\mathbf{F}$, a flat subspace $\mathbf{N}$ and a null space $\mathbf{O}$ and regularizes eigenvalues differently in these three subspaces. The median operation well separates the highly disparate eigenvalues from the stabilized ones. As eigenvalues tend toward equality in $\mathbf{N}$ and are highly disparate in $\mathbf{F}$, the distance $d_{m,r} = \lambda_{med} - \lambda_r$ that indicates the half range of the eigenvalue variation in $\mathbf{N}$ is very small comparing to the highly disparate eigenvalues in $\mathbf{F}$. Therefore, $\lambda_{med} + d_{m,r}$ is proposed to separate the highly disparate subspace $\mathbf{F}$ from the flat subspace $\mathbf{N}$.

Although the finite number of training samples may also bias the eigenvalues in $\mathbf{F}$, we ignore this very small bias because the population eigenvalues are highly disparate in this subspace. In $\mathbf{N}$, however, the population eigenvalues tend towards equality and hence the finite number of training samples results in much faster decay of the eigenvalues. Therefore, we slow down the decay of the eigenvalues by replacing the real eigenspectrum in $\mathbf{N}$ with the proposed model (4.5) that approximates to the real eigenspectrum in $\mathbf{F}$.

The zeros eigenvalues in the null space is only an estimate on one set of the training data. Adding new training samples will make them nonzero. However, based on the available training data that result in zero eigenvalues, the regularized eigenvalues in the null space should be smaller than those in the other subspaces. As we have no information about the variation of the eigenvalues in the null space, they are regularized as a positive constant smaller than the smallest regularized eigenvalue in the other subspaces.

ML can be seen as another eigenspectrum regularization approach that also keeps eigenvalues in the principal subspace unchanged and only modifies eigenvalues in the complementary subspace. DSL adopts the same regularization method. The both approaches replace eigenvalues in the complementary subspace with the constant calculated by the average eigenvalue over this subspace. Obviously, this will amplify some eigenvalues and attenuate the others in the complementary subspace. The unduely reduced eigenvalues in the complementary subspace may introduce additional over-fitting problem because the differences between these eigenvalues and those in the principal subspace are enlarged.

4.4.3 Computational Complexity

The computational complexity of the proposed ERE method for training is greater than other subspace methods as it evaluates the discriminant value in the full space of $\mathbf{S}^w$. Although the null spaces of $\mathbf{S}^w$ and $\mathbf{S}^b$ both contain discriminative information and hence are preserved in the proposed ERE approach, there is no statistical basis that the intersection of the null spaces of $\mathbf{S}^w$ and $\mathbf{S}^b$ that is the null space of $\mathbf{S}^t$ contributes to the discriminative ability [101]. Therefore, we can apply PCA on $\mathbf{S}^t$ to remove the null space of $\mathbf{S}^t$ first and then apply the ERE approach on the $(l - 1)$ -dimensional subspace. This makes the ERE approach tractable for very large image sizes ($n \gg l$). In the practical applications, training is usually an off-line process and recognition is usually an on-line process. Thus, the recognition time is usually much more critical than the training time. Although the recognition time of the ERE approach is same as other approaches for the same number of features, it can be faster than other approaches for the same recognition rate because the proposed ERE approach, as we will see in the experiments, achieves a given recognition rate with fewer features than other approaches.

4.5 Experiments

In all experiments reported in this chapter, images are preprocessed following the CSU Face Identification Evaluation System [73]. Five databases: ORL, AR, GT and two from FERET are used for testing. Each database is partitioned into training and testing sets. For FERET databases, there is no overlap in subjects between the training and testing sets. As ORL, AR and GT databases have only a small number of subjects, both training and testing sets contain all subjects. However, there is no overlap in the sample image between the training and testing sets. The recognition error rate given in this chapter is the percentage of the incorrect top 1 match on the testing set. The proposed ERE method is tested and compared with PCA with Euclidian distance (PCE), PCA with Mahalanobis distance (PCAM), FDA, ML, DSL and UFS approaches in four figures. Furthermore, three experimental results are numerically recorded in two tables, which include more results for comparison, such as results of ERE using $\tilde{\mathbf{S}}^b$ (ERE- $\tilde{\mathbf{S}}^b$), NLDA and SRLDA (a simple regularized LDA that replaces all zero eigenvalues of $\mathbf{S}^w$ with the minimum nonzero eigenvalue). The parameters of UFS are applied that result in the best performance through an exhaustive search in the experiments of [13]. We implement FDA by applying PCA to reduce the dimensionality to $\eta(l - p)$, $0 < \eta \leq 1$ because the rank of $\mathbf{S}^w$ is $l - p$ at most. We present the best result of FDA with η varying from 0.7 to 1.

4.5.1 Results on FERET Database 1

In FERET database, the face image variations include facial expression and other details (like glasses or no glasses), illumination, pose, and aging [45]. We select 2388 images comprising of 1194 subjects (two images per subject) from this database. Some of the sample images are displayed in Fig. 4.4. For this database, images are cropped into the size of 38×33 . In the first experiment, images of 250 people are randomly selected for training and the remaining images of 944 people are used

Figure 4.4: Sample images from the FERET database for four subjects with two sample images per subject.

for testing. Fig. 4.5 shows the recognition error rate on the testing set against the number of features d used in the matching. Note that for ML approach, the number of features used in the matching is constantly the image dimensionality n rather than d .

Figure 4.5: Recognition error rate against the number of features used in the matching on the FERET database of 500 training images (250 people) and 1888 testing images (944 people).

Fig. 4.5 shows that ML outperforms FDA for small d because ML has a large average eigenvalue ρ in the complementary subspace while FDA suffers over-fitting problem due to the small eigenvalues of $\mathbf{S}^w$. However, for large d , the small ρ results in more over-fitting of ML than FDA because ML uses the inverse of ρ to scale features of dimensions from $d + 1$ to n while FDA discards dimensions

from $\eta(l - p) + 1$ to n . UFS outperforms ML and FDA as it solves the over-fitting problem well by reducing the dimensionality to 150. DSL outperforms ML, FDA and UFS when larger number of features are applied. This shows that the null space indeed contains useful discriminative information. However, extracting discriminative features separately from the two subspaces is not efficient for small number of features ($d < 40$ in Fig. 4.5). The proposed ERE approach consistently outperforms other approaches for all number of features and the accuracy gain is significant for smaller number of features.

In the second experiment, more training samples (497 people) are randomly selected and the remaining images of 697 people are used for testing. This database is constructed similar to that of [13]. The recognition error rates are recorded in Table 4.1. All approaches show lower recognition error rates than those in Fig. 4.5 due to the larger number of training samples. The relative performances among ML, FDA and DSL are similar to those in Fig. 4.5. However, UFS achieves higher accuracy gain than ML, FDA and DSL. The ERE approach again consistently outperforms other approaches for all number of features. The small differences of the ERE approaches using $\tilde{\mathbf{S}}^t$ and $\tilde{\mathbf{S}}^b$ are also shown in Table 4.1.

Table 4.1: Recognition Error Rate of Different Approaches for Different Number of Features.

Database	FERET (994/1394 training/testing images)							ORL (leave-one-out training-testing)						
# Feature	20	30	40	50	80	110	140	10	12	14	16	22	30	38
PCAE	35.87	31.13	28.26	27.26	24.82	23.39	22.24	11.00	9.00	9.00	8.25	7.00	7.00	7.25
PCAM	42.18	30.85	23.24	20.52	16.79	16.79	17.36	12.00	8.75	8.50	8.25	7.00	5.00	5.00
FDA	14.35	10.76	8.75	9.33	8.46	8.61	8.18	6.25	5.25	3.75	3.00	2.50	2.50	2.00
ML	7.32	7.03	7.75	7.60	8.46	8.75	8.90	3.50	3.50	3.00	2.25	2.75	2.75	2.50
DSL	10.90	6.74	6.03	4.73	3.87	4.16	3.87	14.00	9.75	5.75	5.00	3.00	3.00	2.00
UFS	6.60	5.17	4.02	3.59	4.02	4.30	4.45	4.50	3.50	3.25	3.50	2.00	1.75	1.25
NLDA	10.62	8.46	7.32	7.32	6.31	6.89	7.17	6.00	4.25	3.00	3.00	2.50	2.25	2.00
SRLDA	8.46	7.03	5.88	5.74	4.88	5.31	5.88	4.50	3.25	3.25	2.25	2.00	2.00	1.25
ERE_ $\tilde{\mathbf{S}}^b$	5.31	4.45	3.01	2.87	3.30	3.44	3.30	2.50	2.50	2.00	1.25	1.50	1.00	0.75
ERE_ $\tilde{\mathbf{S}}^t$	5.17	4.30	3.01	2.87	3.30	3.44	3.16	2.75	2.25	2.00	1.25	1.50	1.25	1.00

4.5.2 Results on FERET Database 2

This database is constructed, similar to one data set used in [110], by choosing 256 subjects with at least four images per subject. However, we use the same number of images (four) per subject for all subjects. 512 images of the first 128 subjects are used for training and the remaining 512 images serve as testing images. The size of the normalized image is 150×130 , same as that in [110]. For such a large image size, we first apply PCA to remove the null space of $\mathbf{S}^t$ and then apply the ERE approach on the 511-dimensional feature vectors. The training time of the ERE approach by a Matlab program is about 1.3 times of that of FDA approach. The i^{th} images of all testing subjects are chosen to form gallery set and the remaining 3 images per subject serve as the probe images to be identified from the gallery set. Fig. 4.6 shows the average recognition error rates over the 4 probe sets, each of which has a distinct gallery set ($i = 1, 2, 3, 4$).

Figure 4.6: Recognition error rate against the number of features used in the matching on the FERET database of 512 training images (128 people) and 512 testing images (128 people).

Comparing to Fig. 4.5, the recognition error rates of all methods in Fig. 4.6 increase due to larger variation of the testing images. They are also higher than

those in [110] on a similar database because of the different training and testing procedures. In our experiment, there is no overlap in subjects between the training and testing sets and only one image per subject in the gallery set. As the large image size produces a large null space, the variation of the testing images in this large null space results in higher recognition error of ML that uses the whole null space than FDA that discards the null space. Similar to the first two experiments, the proposed ERE approach achieves the lowest recognition error rate for all number of features in Fig. 4.6.

4.5.3 Results on ORL Database

In the experiments on ORL database [111], images are cropped into the size of 57×50 . The ORL database contains 400 images of 40 subjects (10 images per subject). Some images were captured at different times and have different variations including expression (open or closed eyes, smiling or nonsmiling) and facial details (glasses or no glasses). The images were taken with a tolerance for some tilting and rotation of the faces up to 20 degrees. Some of the samples images are displayed in Fig. 4.7.

Figure 4.7: Sample images from the ORL database for two subjects with eight sample images per subject.

We first test various approaches using the first 5 samples per subject for training and the remaining 5 samples per subject for testing. Hence, there are 200 images in the training set and 200 images in the testing set. Fig. 4.8 shows the recognition

error rate on the testing set against the number of features.

Figure 4.8: Recognition error rate against the number of features used in the matching on the ORL database of 200 training images (40 people) and 200 testing images (40 people).

As the training set has only 200 images, it does not well represent the variations of testing images. Therefore, the small principal space does not capture the discriminative information well. This results in poor performance of FDA. UFS discards more dimensions and hence performs worse than FDA. DSL that extracts features in two complementary subspaces is better than FDA. ML that works in the whole space is better than DSL. Again, the proposed ERE approach consistently outperforms other approaches for all number of features.

As ORL database is small, we conduct another experiment with leave-one-out training and testing strategy. In each of the 400 runs of training and testing, one sample is picked out for testing and the remaining 399 samples are included in the training set. The testing results are numerically recorded in Table 4.1. As all images but one are used in training, a small principal subspace on the training data can well represent the testing images. This results in better performances of FDA and UFS that only use the principal subspace than DSL and ML that use both the

principal and its complementary subspaces. However, the proposed ERE approach that works on the entire space outperforms FDA and UFS. It shows that, for this training task, the complementary subspace is still useful but not well handled by the DSL and ML algorithms.

4.5.4 Results on AR Database

AR database has frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). The pictures were taken under no restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. Each subject participated in two sessions, separated by two weeks (14 days) time. The same pictures were taken in both sessions [77]. Some of the sample images are displayed in Fig. 4.9. The color images from this database are converted to gray-scale and cropped into the size of 170×120 , same as the image size used in [77, 112]. 50 subjects with 12 images of frontal illumination per subject were used in [77] and the same amount of subjects with 14 non-occluded images per subject were used in [14]. In our experiment, 75 subjects with 14 non-occluded images per subject were selected from the AR database. The first 7 images of all subjects are used in the training and the remaining 7 images serve as testing images. For this large image size, we first apply PCA to remove the null space of $\mathbf{S}^t$ and then apply the ERE approach on the 524-dimensional feature vectors. The training time of the ERE approach by a Matlab program is about 1.4 times of that of FDA approach. Fig. 4.10 shows the recognition error rate.

As images of AR database were taken under tightly controlled conditions of illumination and viewpoint [77], most approaches achieve similar results. The training set seems to represent the test set very well so that the FDA, UFS and PCAM approaches that only use the principal subspace slightly outperform ML and DSL that use both the principal and its complementary subspaces. Furthermore, PCAM that uses leading eigenvectors of $\mathbf{S}^t$ surprisingly outperforms the LDA based FDA, DSL and UFS approaches. It shows that the problems of small eigen-

Figure 4.9: Sample images from the AR database for one subject with fourteen sample images.

Figure 4.10: Recognition error rate against the number of features used in the matching on the AR database of 525 training images (75 people) and 525 testing images (75 people).

values of $\mathbf{S}^w$ dominate other factors for this tightly controlled database. Although the proposed ERE approach evaluates the discriminant value in the whole space of $\mathbf{S}^w$, it consistently outperforms all other approaches for all number of features. This demonstrates the effectiveness of the proposed eigenspectrum regularization method in alleviating the over-fitting problem.

4.5.5 Results on Georgia Tech Database

The Georgia Tech (GT) Face Database [113] consists 750 color images of 50 subjects (15 images per subject). Some of the GT sample images are shown in Fig. 4.11.

Figure 4.11: Sample images from the GT database for two subjects with eight sample images per subject.

For most of the subjects the face images were taken in two or three sessions over a period of three months, allowing for strong variation in size, facial expression, illumination, and rotation in both the image plane and perpendicular to the image plane. These images are converted to gray-scale and cropped into the size of 112×92 . The first 8 images of all subjects are used in the training and the remaining 7 images serve as testing images. The testing results are numerically recorded in Table 4.2. For this database that has large variations in both pose and expression and some illumination changes, Table 4.2 shows that the FDA, ML, NLDA, SRLDA, DSL and UFS methods have similar performances, which are significantly better than PCAE and PCAE. The proposed ERE approaches, ERE- $\tilde{\mathbf{S}}^b$ and ERE- $\tilde{\mathbf{S}}^t$ both consistently outperform all other approaches for all

Table 4.2: Recognition Error Rate of Different Approaches for Different Number of Features.

Database	GT (400 / 350 training / testing images)										
# Feature	8	12	16	20	24	28	32	36	40	44	48
PCAE	53.71	44.57	37.43	34.57	30.00	29.14	27.71	26.00	26.86	26.86	26.57
PCAM	48.57	38.57	31.43	26.57	26.86	22.29	20.57	20.00	19.71	21.43	19.43
FDA	20.29	16.86	15.43	14.00	12.57	10.29	10.29	9.43	9.43	8.57	8.29
ML	13.71	14.57	12.57	14.29	13.14	13.43	13.14	12.57	12.86	12.86	14.00
DSL	29.43	22.57	16.00	14.00	14.00	11.71	10.57	11.71	10.86	10.00	9.43
UFS	21.71	15.71	14.29	12.00	11.71	10.86	10.00	9.43	9.71	9.14	9.14
NLDA	20.57	13.14	12.57	12.86	11.43	11.14	11.43	11.71	12.57	12.86	11.71
SRLDA	19.14	11.71	11.43	11.71	11.71	10.29	11.14	10.57	11.43	11.14	9.71
ERE- $\tilde{\mathbf{S}}^b$	17.14	10.57	8.57	7.14	7.43	8.26	8.29	7.43	8.00	7.71	7.71
ERE- $\tilde{\mathbf{S}}^t$	17.14	10.57	8.29	6.86	7.43	8.57	8.57	7.71	8.29	7.71	7.43

number of features.

4.5.6 Significance of the Proposed Approach

We have performed 7 sets of experiments with 5 different databases. The proposed ERE approach shows superior performance in the following 5 aspects: First, ERE consistently outperforms all other approaches in all 7 experiments, while no other approach can perform the second best consistently in all experiments. Second, ERE consistently outperforms all other approaches for all number of features (note that ML uses n rather than d features in the matching process). In contrast to that, no other approach can perform the second best for all number of features even in a single experiment. Third, ERE achieves the best performance in all experiments without tuning its parameter. ERE has only one free parameter μ in (4.8). Choosing some proper values of μ in different experiments will further enhance its recognition performances. Fourth, although ERE outperforms certain other approaches only marginally for certain number of features in certain experiments, significant better performances of the ERE approach comparing to these approaches can always be found in at least three other experiments. Fifth, ERE significantly outperforms all other approaches for small number of features.

This demonstrates that the proposed ERE approach extracts more discriminative features than others.

4.6 Summary

This chapter addresses problems of extracting discriminant eigenfeatures from the face image based on a set of training samples. Noise disturbance and finite number of training samples in practice may cause eigenvalues of a scatter matrix deviating from the true variances of the images projected on the corresponding eigenvectors. This deviation may result in recognition performance deterioration for various subspace approaches that scale the eigenfeature by the inverse of the square root of the eigenvalue. Especially for small eigenvalues, their inverses are highly sensitive to the noise disturbance and the effect of the finite number of training samples. A different training set may easily change these values substantially. Therefore, for a good generalization, these small and zero eigenvalues should be regularized. Another problem addressed in this chapter is the null space of the scatter matrix. Both the principal and the null spaces of the within-class scatter matrix contain discriminative information. Neither of them should be simply discarded in the feature extraction process. An optimal feature vector may reside in dimensions that lies in the both subspaces, in other words, it may have nonzero components in the both subspaces. It is inefficient or only suboptimal to construct a feature vector by extracting features separately from the principal and the null spaces. The discriminant evaluation in the whole space leads to a more efficient feature representation of a face image.

In this chapter, eigenspace spanned by the eigenvectors of the within-class scatter matrix is decomposed into a reliable subspace, an unstable subspace and a null subspace. Eigenfeatures are regularized differently in these three subspaces based on an eigenspectrum model to alleviate problems of instability, over-fitting or poor generalization. The discriminant evaluation is performed in the whole space and

the feature extraction or dimensionality reduction is done only at the final stage after the discriminant assessment. This facilitates a discriminative and stable low-dimensional feature representation of the face image. Extensive experiments on the FERET, ORL, AR and GT databases with different numbers of training samples demonstrate that the proposed approach consistently outperforms Eigenface, Fisherface, Bayesian maximum likelihood, null space LDA, dual-space LDA and unified subspace framework. Especially, it achieves more accuracy gains for smaller number of features and for smaller size of training set. This verifies that the proposed approach is more efficient in the feature extraction and more stable, less over-fitting the training data or better generalization.

Chapter 5

Complete Discriminant Evaluation and Feature Extraction for Face Recognition in Kernel Space

5.1 Introduction

Face recognition has drawn considerable attention in biometric society because of its property of non-intrusiveness, which is recognizing a person from a distance. Numerous linear and nonlinear subspace based face recognition methods have been proposed in the last two decades [3, 7, 6]. The most popular nonlinear methods are kernel principal component analysis (KPCA) [28] and kernel Fisher discriminant analysis (KFDA) [42]. These kernel based methods and their variations encode pattern information based on higher order dependencies and are tactful to the dependencies among pixels in the samples. The kernel mappings can capture the nonlinearities and complex relationships among the input data that exist due to the expression, illumination and pose variations.

The basic idea of these kernel methods is to apply nonlinear mapping $\Theta : X \in \mathbb{R}^n \rightarrow \Theta(X) \in \mathbb{H}$ in the image space $\mathbb{R}^n$, followed by linear subspace methods like PCA and FDA in the mapped feature space $\mathbb{H}$. Examples include KPCA [28] and KFDA [42, 114]. Since the feature space $\mathbb{H}$ can be very high or possibly infinite dimensional and the orthogonality needs to be characterized in such a space, it is reasonable to view $\mathbb{H}$ as a Hilbert space. It is difficult to compute the dot products in the high dimensional feature space $\mathbb{H}$. Instead of mapping the data explicitly, the feature space can be computed by using the kernel trick, in which the inner products $\langle \Theta(X_{ij}), \Theta(X_{st}) \rangle$ in $\mathbb{H}$ can be replaced with a kernel function $K(X_{ij}, X_{st})$, where $K(X_{ij}, X_{st}) = \langle \Theta(X_{ij}), \Theta(X_{st}) \rangle$ and X_{ij}, X_{st} are sample vectors in the image space $\mathbb{R}^n$. So, the nonlinear mapping Θ can be performed implicitly in image space $\mathbb{R}^n$ [64, 29]. Numerous studies [6, 115, 116] demonstrate that these kernel based approaches are effective in some real-world applications. However, the basic subspace analysis has still outstanding challenging problems when applied to the face recognition due to the high dimensionality of the face image and the finite number of training samples in practice. Although, in this chapter, we take advantages from the nonlinear or kernel mapping, it does not explore the optimization of kernel mapping functions or any of its parameters. Nevertheless, throughout this chapter, we assume a popular nonlinear mapping function with its parameters fixed for all the experiments and perform our proposed algorithm in this nonlinearly mapped feature space.

Most of the kernel subspace based face recognition methods perform dimensionality reduction or discard a subspace before the discriminant evaluation. A popular method called kernel Fisherface [116] applies PCA first for dimensionality reduction so as to make the within-class scatter matrix nonsingular before the application of LDA. However, applying PCA for dimensionality reduction may lose important discriminative information [14, 100, 88, 38]. The null space approach, NKDA [104] eliminates the principal subspace and extracts eigenfeatures only from the eigenvectors corresponding to the zero eigenvalues. Therefore, NKDA assumes that the null space contains the most discriminative information which is contradictory to

KFDA.

Kernel Direct-LDA (KDDA) method [117] first removes the null space of the between-class scatter matrix and then extracts the eigenvectors corresponding to the smallest eigenvalues of the within-class scatter matrix. It is an open question of how to scale the extracted features as the smallest eigenvalues are very sensitive to noise. A common problem of KFDA, NKDA and KDDA approaches is that they all lose some discriminative information, either in the principal or in the null space because they perform the discriminant evaluation in a subspace.

In fact, the discriminative information resides in both subspaces. Recently, Yang *et al.* [95] proposed a complete kernel Fisher discriminant framework (CKFD), where features extracted from the two complementary subspaces are combined by a summed distance measures in the recognition phase [95]. Dual-space based LDA approach is proposed in [88], where features are scaled in the complementary subspace by an average eigenvalue of the within-class scatter matrix over this subspace. As eigenvalues in this subspace are not well estimated, their average may not be a good scaling factor relative to those in the principal subspace. Open questions of these approaches are how to divide the space into the principal and the complementary subspaces and how to apportion a given number of features to the two subspaces. Furthermore, as the discriminative information resides in the both subspaces, it is inefficient and only suboptimal to extract features separately from the two subspaces.

In this chapter, we propose a method which utilizes the ratios of the successive eigenvalues of the eigenspectrum (shown in Fig. 5.1) to decompose the kernel within-class eigenspace into two subspaces: a reliable subspace spanned mainly by the facial variation and an unreliable subspace due to limited number of training samples. A weighting function (shown in Fig. 5.2) is proposed which circumvents undue scaling of projection vectors corresponding to the unreliable small and zero eigenvalues. Finally, features are extracted based on the discriminant evaluation in the whole kernel eigenspace. In the next section, we first study the behavior of

the unreliable small eigenvalues of within-class variation matrix, then propose a methodology to decompose the eigenspace into principal and unreliable subspaces. Eigenfeature scaling and extraction are presented in section 5.3. Experimental results and discussions are presented in section 5.4 before drawing summary in section 5.5.

5.2 Kernel Feature Scaling and Subspace Decomposition

5.2.1 Overview of Kernel Discriminant Analysis

For a nonlinear mapping Θ , the image data space $\mathbb{R}^n$ can be mapped into the feature space $\mathbb{H}$:

$$\Theta : X \in \mathbb{R}^n \rightarrow \Theta(X) \in \mathbb{H}. \quad (5.1)$$

Consequently, a pattern in the original image space $\mathbb{R}^n$ is mapped into a potentially much higher dimensional feature vector in the feature space $\mathbb{H}$. The within-class scatter matrix is defined by

$$\mathbf{S}^w = \frac{1}{p} \sum_{i=1}^p \frac{1}{q_i} \sum_{j=1}^{q_i} (\Theta(X_{ij}) - \overline{\Theta(X_i)}) (\Theta(X_{ij}) - \overline{\Theta(X_i)})^T, \quad (5.2)$$

where $\overline{\Theta(X_i)} = \frac{1}{q_i} \sum_{j=1}^{q_i} \Theta(X_{ij})$. The between-class scatter matrix $\mathbf{S}^b$ is defined by

$$\mathbf{S}^b = \frac{1}{p} \sum_{i=1}^p (\overline{\Theta(X_i)} - \overline{\Theta(X)}) (\overline{\Theta(X_i)} - \overline{\Theta(X)})^T, \quad (5.3)$$

where $\overline{\Theta(X)} = \frac{1}{p} \sum_{i=1}^p \overline{\Theta(X_i)}$, assuming all classes have equal prior probability.

The well known Fisher objective function [4] can be written in the mapped space

$\mathbb{H}$ as

$$J(\mathbf{\Omega}) = \arg \max_{\mathbf{\Omega}} \frac{|\mathbf{\Omega}^T \mathbf{S}^b \mathbf{\Omega}|}{|\mathbf{\Omega}^T \mathbf{S}^w \mathbf{\Omega}|}. \quad (5.4)$$

Because any solution $\mathbf{\Omega} \in \mathbb{H}$ must lie in the span of all the samples in $\mathbb{H}$, there exist coefficients ϕ_{ij} , such that

$$\mathbf{\Omega} = \sum_{i=1}^p \sum_{j=1}^{q_i} \phi_{ij} \Theta(X_{ij}). \quad (5.5)$$

Combining (5.4) and (5.5), we have [93]

$$\mathbf{\Omega}^T \mathbf{S}^w \mathbf{\Omega} = \mathbf{\Phi}^T \mathbf{S}_{\Theta}^w \mathbf{\Phi}, \quad (5.6)$$

$$\mathbf{\Omega}^T \mathbf{S}^b \mathbf{\Omega} = \mathbf{\Phi}^T \mathbf{S}_{\Theta}^b \mathbf{\Phi}, \quad (5.7)$$

where $\mathbf{\Phi} = \{\phi_{ij}\}$ and

$$\begin{cases} \mathbf{S}_{\Theta}^w = \frac{1}{p} \sum_{i=1}^p \frac{1}{q_i} \sum_{j=1}^{q_i} (\zeta_{ij} - \mu_i)(\zeta_{ij} - \mu_i)^T, \\ \zeta_{ij} = (K(X_{11}, X_{ij}), K(X_{12}, X_{ij}), \dots, K(X_{pq_p}, X_{ij}))^T, \end{cases} \quad (5.8)$$

$$\begin{cases} \mu_i = \frac{1}{q_i} \sum_{j=1}^{q_i} \zeta_{ij}, \\ \mathbf{S}_{\Theta}^b = \frac{1}{p(p-1)} \sum_{i=1}^p \sum_{j=1}^p (\mu_i - \mu_j)(\mu_i - \mu_j)^T. \end{cases} \quad (5.9)$$

So the solution of function (5.4) can be obtained by maximizing

$$J(\mathbf{\Phi}) = \arg \max_{\mathbf{\Phi}} \frac{|\mathbf{\Phi}^T \mathbf{S}_{\Theta}^b \mathbf{\Phi}|}{|\mathbf{\Phi}^T \mathbf{S}_{\Theta}^w \mathbf{\Phi}|} \quad (5.10)$$

and the problem of kernel discriminant analysis is converted into finding the leading eigenvectors of $\mathbf{S}_{\Theta}^{w-1} \mathbf{S}_{\Theta}^b$ [94]. However, in practice, the inversion of $\mathbf{S}_{\Theta}^w$ is impossible as it is often singular due to the limited number of training samples.

Let $\mathbf{S}_{\Theta}^g, g \in \{w, b\}$ represent one of the above scatter matrices. If we regard the elements of the kernel vector or the class mean vector as features, these preliminary

features will be de-correlated by solving the eigenvalue problem

$$\mathbf{\Lambda}^g = \mathbf{\Phi}^{gT} \mathbf{S}_{\Theta}^g \mathbf{\Phi}^g, \quad (5.11)$$

where $\mathbf{\Phi}^g = [\phi_1^g, \dots, \phi_l^g]$ is the eigenvector matrix of $\mathbf{S}_{\Theta}^g$, and $\mathbf{\Lambda}^g$ is the diagonal matrix of eigenvalues $\lambda_1^g, \dots, \lambda_l^g$ corresponding to the eigenvectors. We assume that the eigenvalues are sorted in descending order $\lambda_1^g \geq \dots \geq \lambda_l^g$. The plot of eigenvalues λ_k^g against the index k is called eigenspectrum of the training data in the nonlinear plane. It plays a critical role in the subspace methods as the eigenvalues are used to scale and extract features.

5.2.2 Problems in Feature Scaling and Extraction of KFDA

If we compute all the eigenvalues $\text{diag}(\mathbf{\Lambda}^w) = [\lambda_1^w, \dots, \lambda_l^w]$ and eigenvectors $\mathbf{\Phi}^w = [\phi_1^w, \dots, \phi_l^w]$ of the l -by- l dimensional matrix $\mathbf{S}_{\Theta}^w$ using (5.11), the projection matrix $\bar{\mathbf{\Phi}}^w = [\phi_1^w/\sigma_1^w, \dots, \phi_l^w/\sigma_l^w]$ is so called whitened eigenvector matrix of $\mathbf{S}_{\Theta}^w$ with $\|\phi_k^w\| = 1$ and $\sigma_k^w = \sqrt{\lambda_k^w}$. This implies that if any one of the eigenvalues in (5.11) of these matrices is zero then the corresponding eigenvector (5.10) gets an infinite weighting factor. Since the scatter matrices (5.8) and (5.9) are singular [4, 117, 89], in practice, most of the subspace based algorithms circumvent this problem by ignoring the eigenvectors corresponding to zero eigenvalues. However, as pointed out earlier that the null space of $\mathbf{S}_{\Theta}^w$ contains indispensable discriminative information essential for improving recognition accuracy.

The above argument can be viewed as an l -dimensional pattern vector ζ_{ij} is first represented by an l -dimensional eigenfeature vector $Y_{ij} = \mathbf{\Phi}^{wT} \zeta_{ij}$, and then multiplied by a weighting function

$$w_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq r_w \\ 0, & r_w < k \leq l \end{cases}, \quad (5.12)$$

as shown in Fig. 5.2, where r_w is the rank of $\mathbf{S}_{\Theta}^w$. It is apparent from (5.12) that

the eigenvectors $\{\phi_k^w\}_{k=r_w+1}^l$ or the null space of $\mathbf{S}_\Theta^w$ are weighted by zero and thus the corresponding eigenvectors fail to contribute to the whole space discriminant evaluation, which is done in the later portion of the algorithm. This is unreasonable because features in the null space have zero within-class variances based on the training data and hence should be more heavily weighted. It seems anomalous that the weighting function increases with the decrease of the eigenvalues and then suddenly has a big drop from the maximum value to zero as shown in Fig. 5.2. Furthermore, weights determined by the inverse of σ_k^w is, though optimal in terms of the ML estimation, dangerous when σ_k^w is small ($m < k \leq r_w$). The small and zero eigenvalues are training-set-specific and very sensitive to different training sets. Adding new samples to the training set or using different training set may easily change some zero eigenvalues to nonzero and make some very small eigenvalues several times larger. Therefore, these eigenvalues of the within-class scatter matrix are unreliable.

5.2.3 Eigenratiospectrum and Subspace Decomposition

In order to alleviate the above problem we first work on the eigenspectrum of the within-class variation matrix. It is not difficult to estimate the rank of $\mathbf{S}_\Theta^w$, which is $r_w \leq (l - p)$. The eigenvalues whose indices are close to r_w ($m < k \leq r_w$) are very small or close to zero, their inverses give undue overemphasis to the eigenvectors corresponding to this region (or indices) as shown in Fig. 5.2.

In chapter 4, we regularize eigenvalues in the linear space by modeling the eigenspectrum using a $1/f$ function to fit the most reliable portion of the eigenspectrum. The less reliable portion of the eigenspectrum is then replaced by the model. However, in the kernel space, the model of $1/f$ function cannot fit the kernel eigenspectrum well due to the nonlinear transform of face images. A very clear fact in our experiments is that the eigenspectrum in the kernel space decays much faster than that in the linear space. Thus, in this work, we shall employ different approach to find the unreliable eigenvalues in the kernel space and replace them by a constant

instead of the $1/f$ function.

To differentiate the unreliable eigenvalues from the larger ones we employ the ratios of the successive eigenvalues of the eigenspectrum to decompose the whole eigenspace into two subspaces: a principal or reliable subspace spanned mainly by the facial variation, $\mathbf{P} = \{\phi_k^w\}_{k=1}^m$ and an unreliable or noise dominating subspace due to limited number of training samples, $\bar{\mathbf{P}} = \{\phi_k^w\}_{k=m+1}^l$. For a clearer illustration, we first define the eigenratios as $\mathbf{\Gamma}_\Theta^w = \{\gamma_1^w, \dots, \gamma_{r_w-1}^w\}$, such that

$$\gamma_k^w = \frac{\lambda_k^w}{\lambda_{k+1}^w}, \quad 1 \leq k < r_w. \quad (5.13)$$

The plot of eigenratios γ_k^w of a typical real eigenspectrum against the index k is called kernel eigenratiospectrum of the training data as shown in Fig. 5.1.

Figure 5.1: Eigenratiospectrum (5.13) from a typical real kernel eigenspectrum of $\mathbf{S}_\Theta^w$.

For a robust training, the database size should be significantly larger than the (face or reliable) dimensionality m . We examined several different face databases, the eigenratio plots shown in Fig. 5.1 is a general behavioral pattern that all the eigenratios of different databases portray. It is apparent from the graph that the eigenratios first decreases very rapidly, then stabilizes and finally increases. The increase of the eigenratios should not be the behavior of the true variances

but occurs due to the limited number of training samples. The corresponding eigenvalues are therefore unreliable. Thus, one robust way of finding such a point would be finding the minimum of the eigenratios. The start point of the unreliable region $m + 1$ is estimated by

$$\gamma_{m+1}^w = \min\{\forall \gamma_k^w, \quad 1 \leq k < r_w\}. \quad (5.14)$$

A typical such m value of a real kernel eigenspectrum is shown in Fig. 5.1.

The main purpose of finding the value of m using the eigenratios is to distinguish the reliable eigenvalues from the unreliable ones, which facilitates the decomposition of the entire eigenspace into reliable $\mathbf{P}$ and unreliable $\bar{\mathbf{P}}$ subspaces. Eigenvalues in the unreliable subspace $\bar{\mathbf{P}}$, spanned by $\{\phi_k^w\}_{k=m+1}^l$, will be regularized to facilitate the discriminant evaluation and feature extraction from the whole space of $\mathbf{S}_\Theta^w$ matrix (as described in the next section).

5.3 Scaling of Kernel Eigenvectors and Feature Extraction

5.3.1 Scaling of Kernel Eigenvectors

As pointed out in [107], the largest sample-based eigenvalues are biased high and the smallest ones are biased low due to the finite number of training samples. The eigenspectrum in the principal/reliable subspace is dominated by the face structural component, hence, we keep the eigenvalues in the principal subspace unchanged. In the unreliable subspace $\bar{\mathbf{P}}$, however, the limited number of training samples results in faster decay of the eigenvalues than the true variances. Therefore, the decay of the eigenvalues should be slowed down to compensate the effect of the finite number of training samples.

From Fig. 5.2 it is evident that when the inverses of the eigenvalues $\{\lambda_k^w\}_{k=m+1}^l$

are used for feature weighting (5.12), the corresponding eigenvectors get undue over-scaling in this range. We should not trust the eigenvalues, $\lambda_k, k > m$ as they are greatly effected by the finite number of training samples. Moghaddam *et al.* [44] replaces the small and zero eigenvalues by the average eigenvalue over the unreliable subspace and Wang *et al.* [88] adopted it in the dual-space approach. However, this may introduce additional over-fitting problem in dimensions of the unreliable subspace whose eigenvalues are larger than the average. (There must be some eigenvalues in the unreliable subspace larger than the average being replaced by the smaller average eigenvalue). As eigenvalues in this subspace are biased smaller, their average may not be a good scaling factor relative to those in the principal subspace. Therefore, we propose to replace the unreliable eigenvalues $\{\lambda_k^w\}_{k=m+1}^l$ by the upper bound eigenvalue of the unreliable subspace, i.e.

$$\lambda_{const}^w = \max\{\forall \lambda_k^w, \quad k \geq m\}. \quad (5.15)$$

Thus, the final weighting function can be written as

$$\tilde{w}_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq m \\ 1/\sqrt{\lambda_{const}^w}, & m < k \leq l \end{cases}. \quad (5.16)$$

Fig. 5.2 shows the proposed feature weighting function $\tilde{w}_k^w$ calculated by (5.13), (5.14), (5.15) and (5.16) comparing with that w_k^w of (5.12). The new weighting function $\tilde{w}_k^w$ is identical to w_k^w in the principal space and remains constant in the unreliable and null subspaces. Note that, different from those in [44, 88], no eigenvalue in the unreliable space becomes smaller in our approach.

Using this weighting function and the eigenvectors ϕ_k^w , training pattern data are transformed to

$$\tilde{Y}_{ij} = \tilde{\Phi}_l^{w^T} \zeta_{ij}, \quad (5.17)$$

Figure 5.2: Weighting functions of (5.12) and (5.16) in the principal- and unreliable-subspaces based on a typical real kernel eigenspectrum.

where

$$\tilde{\Phi}_l^w = [\tilde{w}_k^w \phi_k^w]_{k=1}^l = [\tilde{w}_1^w \phi_1^w, \dots, \tilde{w}_l^w \phi_l^w]. \quad (5.18)$$

There is no dimension reduction in this transformation as $\tilde{Y}_{ij}$ and ζ_{ij} have the same dimensionality l .

Problems of dimensionality reduction of KFDA were also discussed in [94], where a kernel machine-based regularized Fisher discriminant (K1PRFD) algorithm was proposed. This approach regularizes the within-class scatter matrix by adding a constant to all eigenvalues. As pointed out in [107], the bias of eigenvalues is most pronounced when the eigenvalues tend toward equality, and it is much less severe when their values are highly disparate. For the application of face recognition, it is well-known that the eigenspectrum first decays very rapidly and then stabilizes. Hence, adding a constant to the eigenspectrum may bias back the rapidly changing eigenvalues in principal space too much that introduces additional error source, and bias back the flat eigenvalues in null space too little at the same time [107].

5.3.2 Kernel Eigenfeature Extraction

After the feature scaling, a new between-class scatter matrix is formed by vectors $\tilde{Y}_{ij}$ of the training data as

$$\tilde{\mathbf{S}}_{\Theta}^b = \frac{1}{p} \sum_{i=1}^p (\tilde{Y}_i - \bar{Y})(\tilde{Y}_i - \bar{Y})^T, \quad (5.19)$$

where $\tilde{Y}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} \tilde{Y}_{ij}$ and $\bar{Y} = \frac{1}{p} \sum_{i=1}^p \frac{1}{q_i} \sum_{j=1}^{q_i} \tilde{Y}_{ij}$. The weighted features $\tilde{Y}_{ij}$ will be de-correlated for $\tilde{\mathbf{S}}_{\Theta}^b$ by solving the eigenvalue problem as (5.11). Suppose that the eigenvectors in the eigenvector matrix $\tilde{\Phi}_l^b = [\tilde{\phi}_1^b, \dots, \tilde{\phi}_l^b]$ are sorted in descending order of the corresponding eigenvalues. The dimensionality reduction is performed here by keeping the eigenvectors with the d largest eigenvalues

$$\tilde{\Phi}_d^b = [\tilde{\phi}_k^b]_{k=1}^d = [\tilde{\phi}_1^b, \dots, \tilde{\phi}_d^b], \quad (5.20)$$

where d is the number of features usually selected by a specific application. Thus, the proposed feature scaling and extraction matrix $\mathbf{U}_{\Theta}$ is given by

$$\mathbf{U}_{\Theta} = \tilde{\Phi}_l^w \tilde{\Phi}_d^b. \quad (5.21)$$

This transforms a face sample vector ζ_{ij} of dimensionality l into a feature vector F of dimensionality d , by

$$F = \mathbf{U}_{\Theta}^T \zeta_{ij}. \quad (5.22)$$

Note that the dual-space approaches [88, 95] extract features separately from two subspaces. As the discriminative information resides in the whole space, it is inefficient and only suboptimal to extract the discriminative features separately from two subspaces by two separate discriminant evaluations in two subspaces. Our method decomposes the kernel space into two subspaces only for the regularization of the eigenvalues in the unreliable subspace. We extract the discriminative

features from the whole space by searching the most discriminative features in the full space. Thus, our method is based on the global optimization instead of two local optimizations of the dual-space approaches in [88, 95].

5.3.3 The Proposed Algorithm

The proposed complete discriminant evaluation and feature extraction in kernel space for face recognition (CDEFE) approach is summarized below:

At the training stage:

- (1) Given a training set of face image vectors $\{X_{ij}\}$ and a kernel function $K(X_{ij}, X_{st})$, compute ζ_{ij} using (5.8).
- (2) Compute $\mathbf{S}_\Theta^w$ by (5.8) and solve the eigenvalue problem as (5.11).
- (3) Decompose the kernel eigenspace into principal- and unreliable-spaces by determining the m value using (5.13) and (5.14).
- (4) Transform the training pattern samples represented by ζ_{ij} into $\tilde{Y}_{ij}$ by (5.17) with the weighting function (5.16) determined by (5.13), (5.14) and (5.15).
- (5) Compute $\tilde{\mathbf{S}}_\Theta^b$ by (5.19) with $\tilde{Y}_{ij}$ and solve the eigenvalue problem as (5.11).
- (6) Obtain the final feature scaling and extraction matrix by (5.18), (5.20) and (5.21) with a predefined number of features d . ■

At the recognition stage:

- (1) Transform each n -D face image vector X into l -D feature pattern vector ζ_{ij} using the kernel function K and (5.8).
- (2) Transform each l -D feature pattern vector ζ_{ij} into d -D feature vector F by (5.22) using the feature regularization and extraction matrix $\mathbf{U}_\Theta$ obtained in the training stage.

- (3) Apply a classifier trained on the gallery set to recognize the probe feature vectors. (A simple first nearest neighborhood classifier is applied in the experiments.) ■

5.3.4 Computational Complexity

The difference in the computational complexity among various kernel based approaches largely depends on the training procedure and database structure. In general, the computational complexity of the training procedure of the proposed CDEFE approach is $\mathcal{O}(l^3)$, where l is the number of training samples. This is same as other kernel based subspace methods such as KFDA, KDDA, NKDA, CKFD and K1PRFD. However, some approaches require three eigen-decompositions and some need only two. The eigen-decomposition is the most time consuming part in the training. The proposed CDEFE approach requires only two eigen-decompositions. Different from KFDA, KDDA and NKDA methods, however, CDEFE, CKFD, K1PRFD and the dual space dual-space approach [88] need to compute all eigenvectors of the within-class scatter matrix. In this respect, CDEFE, CKFD, K1PRFD and dual-space approach in general may require additional computation comparing to KFDA, KDDA and NKDA. Note that K1PRFD uses an iterative conjugate gradient method to process the eigenvalues, which is time consuming.

After the processing of the within-class scatter matrix, CKFD and dual-space approach require two eigen-decompositions in two complementary subspaces while CDEFE and K1PRFD compute one eigen decomposition in the whole space. As the number of subjects p is at most half of the total number of the training samples l (single sample per subject cannot contribute to the within-class scatter matrix), snapshot method [118, 4] can be used. Using this method, CDEFE only requires one eigen-decomposition of the between-class scatter matrix of dimensionality $p \times p$, which is smaller than that of at least one of the two between-class scatter matrices of CKFD and dual-space approach. In this respect, the proposed CDEFE may require less computation than CKFD and dual-space approach, depending on the

database structure.

The testing/recognition time of KFDA, KDDA, NKDA, K1PRFD and CDEFE are almost same for the same number of features because the recognition procedure is computing a distance between the probe image and all the gallery images. Recognition time of CKFD is slightly more than the above five methods because it uses summed normalized distance between the probe and gallery images [95]. In practical face recognition systems, training is usually an offline process and recognition is usually an online process. Thus, the recognition time is usually more critical than the training time. Although the recognition time of the CDEFE approach is same as other approaches for the same number of features, it can be faster than other approaches for the same recognition rate because the proposed CDEFE approach, as we will see in the experiments, achieves a given recognition rate with fewer features than other approaches.

5.4 Experiments and Discussions

In all experiments reported in this work, images are aligned and normalized following the CSU Face Identification Evaluation System [73]. Four databases: ORL, GT and two from FERET are used for testing. Each database is partitioned into training and testing sets. For FERET databases, there is no overlap in subject between the training and testing sets. As ORL and GT databases have only a small number of subjects, both training and testing sets contain all subjects. However, there is no overlap in the sample image between the training and testg sets. In our experiments, polynomial cosine kernel function is chosen, $K(X_{ij}, X_{st}) = \frac{\tilde{K}(X_{ij}, X_{st})}{\sqrt{\tilde{K}(X_{ij}, X_{ij})\tilde{K}(X_{st}, X_{st})}}$, where $\tilde{K}(X_{ij}, X_{st}) = \langle \Theta(X_{ij}), \Theta(X_{st}) \rangle = (a\langle X_{ij} \cdot X_{st} \rangle + b)^c$, since cosine kernel gave good performances in the experiments of [93, 104] and better performances than the original polynomial kernels [119, 120, 121, 93]. The kernel parameters are set same as that of in [93, 104, 17]. The recognition error rate given in this work is the percentage of the incorrect top 1 match on the testing set.

The proposed CDEFE method is tested and compared with KFDA [42], KDDA [117], NKDA [104], CKFD [95] and K1PRFD [94] approaches. The parameters of CKFD are applied that are mentioned in the experiments of [95].

5.4.1 Results on FERET Database 1

2388 images comprising of 1194 persons (two images FA/FB per person) are selected from the FERET database [45]. Images are cropped into the size of 38×33 . Before comparing the proposed CDEFE method with other state-of-the-art kernel based approaches we first test the three contributions of this work: the subspace decomposition point m , the eigenspectrum regularization and the discriminant evaluation in the whole kernel space. 500 images of 250 subjects are randomly selected for training and the remaining 1888 images of 944 subjects are used for testing. Fig. 5.3 shows the recognition error rate on the testing set against the number of features d used in the matching. In Fig. 5.3, CDEFE-95% represents a

Figure 5.3: Recognition error rate against the number of features used in the matching on the FERET database of 500 training images (250 subjects) and 1888 testing images (944 subjects).

variant of our approach where the subspace decomposition point m is determined

by keeping 95% eigenvalue energy in the reliable subspace instead of using (5.14); CDEFE-Av represents another variant of our approach by replacing λ_{const}^w in (5.16) with the average eigenvalue over the unreliable subspace as used in [44] and in [88]; and CDEFE-Dual represents the third variant of our approach where features are extracted separately from the principal and unreliable subspaces as the idea of the dual-space approach in [88].

Fig. 5.3 shows that our subspace decomposition point is much better than the empirical setting of keeping 95% eigenvalue energy. It also illustrates the advantage of the upper bound eigenvalue instead of the average eigenvalue for the eigenspectrum regularization. Furthermore, Fig. 5.3 clearly demonstrates that the global optimization in the whole space outperforms the two local optimization processes in two subspaces. In fact, the recent proposed CKFD approach [95] also extracts features separately from two subspaces, whose recognition performance is shown in all of the following experiments.

Fig. 5.4 compares the proposed CDEFE method with other state-of-the-art kernel based approaches. Both KFDA and KDDA perform badly because they extract features only from the principal subspace. K1PRFD achieves slightly higher accuracy than NKDA and CKFD as it extracts features from the whole space. Our proposed CDEFE approach consistently outperforms all other approaches for all number of features and the accuracy gain is significant for smaller number of features.

Since this database is larger than others, we conduct a second experiment with larger number of training samples where 497 subjects are randomly selected for training and the remaining images of 697 subjects are used for testing. Fig. 5.5 shows the results. CKFD that uses information from both the subspaces performs better than KFDA, KDDA, NKDA and K1PRFD approaches. Although K1PRFD uses the whole space, it does not outperform CKFD. Possible reasons for this are, as stated previously, that adding a constant to the eigenspectrum may bias back the rapidly changing eigenvalues too much that introduces additional error source, and

Figure 5.4: Recognition error rate against the number of features used in the matching on the FERET database of 500 training images (250 subjects) and 1888 testing images (944 subjects).

bias back the flat eigenvalues too little at the same time. The proposed CDEFE approach again consistently outperforms other approaches.

Figure 5.5: Recognition error rate against the number of features used in the matching on the FERET database of 994 training images (497 subjects) and 1394 testing images (697 subjects).

5.4.2 Results on FERET Database 2

This database is constructed by choosing 256 subjects with at least four images per subject. We use the same number of images (four) per subject for all subjects. 512 images of the first 128 subjects are used for training and the remaining 512 images serve as testing images. The size of the normalized image is 150×130 , same as that in [110]. The i^{th} images of all testing subjects are chosen to form gallery set and the remaining 3 images per subject serve as the probe images to be identified from the gallery set. Therefore, for each run there is only one image per subject in the gallery data set and three images per subject in the testing data set. Fig. 5.6 shows the average recognition error rates over the 4 probe sets, each of which has a distinct gallery set ($i = 1, 2, 3, 4$).

Figure 5.6: Average recognition error rate against the number of features used in the matching on the FERET database of 512 training images (128 subjects) and 512 testing images (128 subjects).

Comparing to Fig. 5.5, the recognition error rates of all methods in Fig. 5.6 increase due to larger variation of the testing images. Unlike the previous two experiments, KDDA performs much better than KFDA because of the availability of more number of samples per subject during training. K1PRFD which uses

the full kernel eigenspace performs better than CKFD, KDDA and KFDA methods. However, K1PRFD does not outperform NKDA consistently. Similar to the first two experiments, the proposed CDEFE approach achieves consistently lowest recognition error rate for all number of features in Fig. 5.6.

5.4.3 Results on ORL Database

Images of the ORL database [111] are cropped into the size of 57×50 . The ORL database contains 400 images of 40 people (10 images per person). In the first experiment, we test various approaches using the first 5 samples per subject (200 images) for training and the remaining 5 samples per subject (200 images) for testing. Fig. 5.7 shows the recognition error rate on the testing set against the number of features. Similar to the previous experiments, our method consistently outperforms others.

Figure 5.7: Recognition error rate against the number of features used in the matching on the ORL database of 200 training images (40 subjects) and 200 testing images (40 subjects).

To obtain more reliable results on ORL database, we conduct another experi-

ment with leave-one-out training and testing strategy. In each of the 400 runs of the training and testing, one sample is picked out for testing and the remaining 399 samples are included in the training set. The testing results are numerically recorded in Table 5.1. KDDA outperforms both KFDA and NKDA but not consistently. Similar to the results of FERET database 2, KDDA performs better when more number of samples per subject are present in the training database. CKFD, which uses a summed normalized distance measures from the two subspaces outperforms KDDA and K1PRFD. However, CDEFE which evaluates the discriminative information in the whole eigenspace consistently outperforms CKFD, K1PRFD and all others. It shows that, for this training task, the complementary subspace is still useful but not well handled by the CKFD and K1PRFD approaches.

Table 5.1: Recognition Error Rate of Different Approaches for Different Number of Features.

Database	ORL (leave-one-out training-testing)							GT (400 / 350 training / testing images)						
# Feature	6	8	10	20	32	36	38	6	10	14	20	28	38	46
KFDA	18.00	9.75	7.25	4.25	2.75	2.75	2.50	38.57	26.57	22.00	18.00	14.86	12.86	12.57
KDDA	15.75	8.50	6.75	2.75	2.50	3.00	2.75	33.14	22.00	15.43	12.00	9.14	9.14	9.43
NKDA	20.75	11.00	8.25	3.50	2.75	2.75	2.75	35.43	20.00	17.14	13.43	13.43	12.29	12.29
CKFD	15.50	9.00	6.00	2.25	2.25	1.75	2.00	38.00	23.43	15.71	12.86	10.86	10.29	8.86
K1PRFD	17.75	10.25	7.00	3.25	3.00	2.50	2.50	31.71	19.43	17.14	13.71	13.14	11.71	12.57
CDEFE	13.25	6.50	5.00	1.75	2.00	1.25	1.25	26.29	17.71	12.86	10.00	8.29	8.29	8.29

5.4.4 Results on Georgia Tech Database

The Georgia Tech (GT) Face Database [113] consists 750 color images of 50 subjects (15 images per subject). These images have large variations in both pose and expression and some illumination changes. Images are converted to gray-scale and cropped into the size of 112×92 . The first 8 images of all the subjects are used in the training and the remaining 7 images serve as testing images. The testing results are numerically recorded in Table 5.1. All the approaches perform relatively similar to the previous experiments. KDDA outperforms CKFD but not consistently. Both KDDA and CKFD outperform KFDA, NKDA and K1PRFD approaches but not

consistently, this probably shows that KDDA and CKFD perform better when more number of samples per subject are present in the training database (similar to the previous experiments). The proposed CDEFE approach again consistently outperforms all other approaches for all number of features.

5.4.5 Summary of Experiments

We have performed 6 sets of experiments with 4 different databases. The proposed CDEFE approach shows superior performance in the following 5 aspects: First, CDEFE consistently outperforms all other approaches in all 6 experiments, while no other approach can perform the second best consistently in all the experiments. Second, CDEFE achieves the lowest recognition error rate consistently for all number of features. In contrast to that, no other approach can perform the second best for all number of features even in a single experiment. Third, CDEFE avoids any kind of heuristic parameter setting in its implementation/application. Fourth, although for certain number of features in certain experiments CDEFE achieves only marginal accuracy gains comparing to some other approaches, significant better performances of the CDEFE approach comparing to these approaches can be found in some other experiments. Fifth, CDEFE significantly outperforms all other approaches for small number of features. This demonstrates that the proposed CDEFE approach extracts more discriminative features than others.

5.4.6 Contributions of the Proposed Approach

There are three contributions in the proposed approach. First, CDEFE uses the minimum point of the eigenratiospectrum to decompose the kernel space into a reliable and an unreliable subspaces, which circumvents the heuristic parameter selection. Second, the proposed approach regularizes the eigenvalues in the unreliable subspace to the constant determined by the upper bound eigenvalue of the unreliable subspace. Different from other approaches that use the average eigen-

value, our method does not diminish any eigenvalue in the unreliable subspace. Third, CDEFE performs the discriminant evaluation in the whole kernel space, thereby, not lose out important discriminative information. Unlike the conventional methods which lose discriminative information either in the principal or in the null space and the recently proposed approaches which extract features separately from the two subspaces, our method extracts features from the whole kernel space which boosts its recognition performance as compared to others. This is verified and demonstrated through extensive experimentations.

5.5 Summary

In this chapter, we have addressed the problems of eigenfeature scaling and its extraction from the whole kernel space. To facilitate the discriminative feature extraction from the whole kernel space, the ratios of the successive eigenvalues in the eigenspectrum are used to decompose the within-class eigenspace into a reliable and an unreliable subspaces. Eigenvalues of the unreliable subspace are regularized to the upper bound eigenvalue of this subspace. This circumvents the undue scaling of the eigenvectors corresponding to the unreliable small and zero eigenvalues and facilitates the feature extraction from the whole kernel space by the discriminant evaluation in the full kernel space. Experiments on the FERET, ORL and GT databases demonstrate that the proposed approach consistently outperforms other popular methods.

Chapter 6

Face Verification Using Eigenmodel

6.1 Introduction

In Biometrics, face recognition has two main applications, one is face verification and the other is face identification. Face verification is a task to determine whether the person claiming a given identity is the true claimant or the person is an imposter. This can be done by computing the distance measures between the probe sample and the samples of the claimed person in the gallery. The final decision is made based on a pre-learned threshold. While for face identification, the system has to determine the identity of a person by computing distance measures between the probe and all the gallery samples in the database. Finally, based on the minimum distance measure or highest similarity score a decision is given. Over the last decade numerous algorithms for face recognition have been proposed [3, 7] and considerable good recognition rates have been achieved. However, these algorithms focus on face identification and very few are actually evaluated for face verification purpose. Many works ignore the intrinsic difference between face verification and face identification tasks [66, 122]. Nevertheless, there are some attempts made to

distinguish face verification from face identification [123, 124].

In general, similar training algorithms can be used for both face verification and identification. However, there are some intrinsic differences between the two cases. While an identification system could be limited to identify an input as one of its known users, a verification system must be able to reject the unknown imposters. Consequently, while the training set could include some sample images of all test subjects for certain identification applications, no sample of test imposters should be included in the training set for a verification system. Hence, a learning algorithm may suffer more severe over-fitting problem (or poor generalization) for face verification task than for the face identification. Furthermore, decisions of a verification system depend on the operating points or thresholds while those of an identification system depend on the rank of the similarity. This leads to quite different evaluation methods for the verification system from the identification system. Thus, a method that has a high accuracy for the identification task may not necessarily achieve a high accuracy for the verification task.

Both face verification and identification are challenging as the human faces can always undergo significant variations in appearance due to facial expressions, illumination changes, shifts or rotations and different pose conditions. In this chapter, we take the advantages of both the proposed eigenfeature regularization scheme and the eigenratiospectrum methodology for the evaluation of face verification process. Two control points, m_1 and m_2 are estimated using (4.8) and (5.14) respectively. The replacement of unreliable small and zero eigenvalues by the modeled values in this range of control points (between m_1 and m_1) and a constant model eigenvalue beyond the eigenspace estimated by (5.14) reduces the sensitivity of discriminative methods to the number of training samples, high dimensionality of the face images and noises in the data. It provides better generalization or less over-fitting as compared to the existing methods for face verification tasks. In addition, many subspace methods are evaluated only for face identification task. Their performances for face verification are unknown in the literature. This work

experimentally evaluates eight popular subspace based approaches for the face verification task and compares them with the proposed method. Experimental results on three face databases demonstrate that the proposed method consistently outperforms others. In the following section, we first discuss the usefulness of our proposed method and then outline the algorithm. Comparisons and analysis are done in section 6.3. Experimental results and discussions that compare our method with others are presented in section 6.4. A summary is drawn in section 6.5.

6.2 Proposed Algorithm

Owing to the fact that a learning algorithm may suffer more severe over-fitting problem (or poor generalization) for face verification task than for the face identification, we propose to restrain the use of the model eigenvalues, the scheme proposed in chapter 4. The advantages of replacing the real eigenvalues with the model eigenvalues are many, however, the eigenvalues corresponding to the indexes (after sorting in descending order) near the rank of $\mathbf{S}^w$ are unreliable because of the limited number of training samples and high dimensionality of face images. Hence, we first estimate two control points: m_1 using (4.8) and m_2 using (5.14). These two control points help us to identify three portions of the eigenspectrum: a reliable face variation dominating subspace $\mathbf{F} = \{\phi_k^w\}_{k=1}^{m_1}$, an unstable noise variation dominating subspace $\mathbf{N} = \{\phi_k^w\}_{k=m_1+1}^{m_2}$ and a null space $\mathbf{O} = \{\phi_k^g\}_{k=m_2+1}^n$ as illustrated in Fig. 6.1 (we keep the same symbols to avoid any further introduction of new symbols).

We propose a methodology of regularization of the eigenspectrum, in which the eigenvalues in the range space between m_1 and m_2 are replaced by model values and a model constant is used to replace the remaining small and zero eigenvalues.

Figure 6.1: Weighting functions of (4.10) and (6.2) in the principal- and unreliable-subspaces based on a typical real eigenspectrum.

Therefore, the proposed regularized eigenspectrum $\hat{\lambda}_k^w$ is given by

$$\hat{\lambda}_k^w = \begin{cases} \lambda_k^w, & k < m_1 \\ \frac{\alpha}{k+\beta}, & m_1 \leq k \leq m_2 \\ \frac{\alpha}{m_2+1+\beta}, & m_2 < k \leq n \end{cases} \quad (6.1)$$

The proposed feature weighting function is then

$$\hat{w}_k^w = \frac{1}{\sqrt{\hat{\lambda}_k^w}}, \quad k = 1, 2, \dots, n. \quad (6.2)$$

This feature weighting function, $\hat{w}_k^w$ is compared with that of w_k^w and are shown in Fig. 6.1. It is noticeable from the figure that the new weighting function $\hat{w}_k^w$ is identical to w_k^w in the face space, increases along with k at a much slower pace than w_k^w in the noise space (between the two control points m_1 and m_2) and has maximal weights instead of zero of w_k^w in the null space (region beyond m_2).

The proposed face verification using modeled eigenspectrum (MES) approach is

summarized below:

At the training stage:

- (1) Given a training set of face image vectors $\{X_{ij}\}$, compute $\mathbf{S}^w$ by using (4.1) and solve the eigenvalue problem as (4.4).
- (2) Determine m_1 using (4.7) and (4.8). Also estimate m_2 using (5.13) and (5.14) determined by the ratios of the eigenvalues (5.13).
- (3) Decompose the eigenspace into face-, noise-, and null-space by using m_1 and m_2 values obtained in step 2.
- (4) Transform the training samples represented by X_{ij} into $\tilde{Y}_{ij}$ using (4.15) with the new weighting function (6.2) (replace $\tilde{w}_k^w$ with $\hat{w}_k^w$ in (4.16)) determined by (4.11), (4.12) and (6.1).
- (5) Compute $\tilde{\mathbf{S}}^t$ by (4.17) with $\tilde{Y}_{ij}$ and solve the eigenvalue problem as (4.4).
- (6) Obtain the final feature scaling and extraction matrix by (4.16), (4.18) and (4.19) with a predefined number of features d . ■

At the enrollment or registration stage:

- (1) Extract d -D feature vector F from the enrolled n -D normalized face image vector X by (4.20) using the feature scaling and extraction matrix $\mathbf{U}$ obtained in the training stage (4.19).
- (2) Store the extracted feature vector and the registration ID into the gallery feature vector set. ■

At the verification stage:

- (1) Extract d -D feature vector F from the enrolled n -D normalized face image vector X by (4.20) using the feature scaling and extraction matrix $\mathbf{U}$ obtained in the training stage (4.19).

- (2) Compare or match the probe feature vector with that in the gallery feature vector set corresponding to the claimed ID. ■

In the experiments of this work, cosine distance measure using (4.21) between a probe feature vector F_P and a gallery feature vector F_G is applied to the proposed approach.

6.3 Analysis and Comparisons

Subspace methods like FDA, DLDA, UFS and NDA approaches discard a subspace before the discriminant evaluation. Therefore, they perform a local optimization process and hence their extracted features are suboptimal or are the most discriminative only in a subspace. Although ML works in the whole space, it does not evaluate the discriminant value and hence the whole face image must be used as features in the verification process. It performs well for the identification problem but as we will see in the experiments, it suffers from severe over-fitting problem for the verification task. Although the dual-space based approaches do not throw away any subspace before the discriminant evaluation, it is inefficient or only suboptimal to evaluate the discriminant value and extract features separately in two subspaces and then to combine them. Other open questions of the dual-space based approaches include how to divide the space into the two subspaces and how to apportion the given number of features to the two subspaces.

The proposed algorithm (MES) uses the three subspace based eigenspectrum decomposition, which differentiates the reliable, unreliable and zero eigenvalues by estimating two control points. Two reasonable decomposition points are determined by the algorithm. This eigenspace decomposition is not used to limit the discriminant evaluation in one subspace but to enable the evaluation in the whole space. Thus, the proposed method is based on the global optimization that extracts features by searching the most discriminative ones in the whole space. The parameters of the eigenspectrum model are determined by the reliable portion of the

real eigenspectrum and the modeled eigenvalues are then used to replace the unreliable small real eigenvalues. Furthermore, the very small and zero real eigenvalues are replaced by a constant model eigenvalue. This scheme reduces the sensitivity of the extracted features to the dimensionality of face image, the number of training samples and noise disturbance. Consequently, the proposed approach alleviates the over-fitting problem that often occurs in the training process with limited number of high dimensional samples. As discussed before, the face verification task may suffer from more severe over-fitting problem (or poorer generalization) than the face identification task. The proposed MES approach facilitates a discriminative and stable low-dimensional feature representation of the face image, which is verified in the following experiments on the face verification task.

6.4 Experimental Results

We evaluate and validate our proposed MES approach on three popular databases: AR database, FERET database 1 and FERET database 2. Each database is partitioned into training and testing sets. In all experiments reported in this work, images are preprocessed following the CSU Face Identification Evaluation System [73]. Face verification is performed by accepting a claimant if the subject's matching (distance) score is less than a threshold and rejecting the claimant if its matching (distance) score is greater than the threshold. Verification performance is evaluated using two measures: false acceptance rate (FAR) and the false rejection rate (FRR). FAR is the ratio of the number of accepted imposter matching to the total number of imposter matching. FRR is the ratio of the number of rejected genuine matching to the total number of genuine matching. The plot of FRR against FAR is called the receiver operating characteristics (ROC) curves. The system performances at various different operating points (thresholds) are characterized by the ROC curve. The equal error rate EER, defined by $EER = FAR = FRR$ at a specific threshold, serves as a single number indicator of a verification systems

performance.

The proposed MES method is tested and compared with PCA [26] with Euclidean distance (PCAE), PCA with Mahalanobis distance (PCAM), FDA [32], NDA [12], DLDA [34], ML [2], UFS [13] and DSL [88] approaches. The proposed MES approach has only one free parameter μ in (4.8). To avoid exhaustive search for the best parameter value, μ is fixed to be 1 in all experiments of this chapter for fair comparisons with other approaches. The parameters of UFS are applied that result in the best performance through an exhaustive search in the experiments of [13]. We conduct the experiments starting with the number of features $d = 10$, incremented by 2 each time up to $p - 1$, where p is the number of training subjects. Experimental results are presented in this chapter for each approach where the minimum EER is obtained.

6.4.1 Results on AR Database

The color images in AR database [77] are converted to gray-scale and cropped into the size of 170×120 , same as the image size used in [77, 112]. The pictures of most subjects were taken in two sessions (separated by two weeks). In our experiment, 75 subjects with 14 non-occluded images per subject were selected from the AR database. The first 7 images of 60 subjects are used in the training and also as gallery images. The remaining 420 images serve as probe genuine images. Rest 15 subjects with 14 images per subject are used as probe imposters. The total number of genuine matches is $7 \times 7 \times 60 = 2,940$ and the total number of imposter matches is $14 \times 7 \times 15 \times 60 = 88,200$. For this large image size, we first apply PCA to remove the null space of $\mathbf{S}^t$ and then apply the MES approach on the 419-dimensional feature vectors. Fig. 6.2 shows the false rejection rate (%) against the false acceptance rate (%).

The ROC curves of PCA does not appear in Fig. 6.2 because its FRRs and FARs are so high that their values are out of the range of Fig. 6.2. Their EERs

Figure 6.2: False rejection rate against the false acceptance rate on the AR database of 420 training images (60 subjects), 420 probe genuine images (60 subjects) and 210 probe imposter images (15 subjects). The total number of genuine matches is $7 \times 60 \times 7 = 2,940$ and the total number of imposter matches is $15 \times 14 \times 60 \times 7 = 88,200$.

are numerically recorded in Table 6.1. We see that ML approach does not perform well for the face verification task although it is one of the best approaches for the face identification task. The problem of ML for face verification was also addressed in [125]. Fig. 6.2 shows that the proposed MES method consistently outperforms all other 8 approaches for all different operating points (thresholds).

6.4.2 Results on FERET Database 1

In FERET database, the face image variations include facial expression and other details (like glasses or no glasses), illumination, pose, and aging [45]. 2388 images comprising of 1194 subjects (two images FA/FB per subject) are selected from the FERET database. Images are cropped into the size of 38×33 . In the first experiment, images of 497 subjects are randomly selected for training and the remaining images of 697 subjects are used for testing. For this database, the subjects which are used for training are different from the testing, there is no overlap in subjects between the training and the testing data sets. The gallery data set contains 697 subjects with 1 image per subject. During testing, 697 subjects are once used as probe genuine and again as probe imposters. The total number of genuine matches is $697 \times 1 = 697$ and the total number of imposter matches is $697 \times 696 = 4,85,112$. Fig. 6.3 shows the false rejection rate (%) against the false acceptance rate (%).

The ROC curves of FDA and DLDA do not appear in Fig. 6.3 because their FRRs and FARs are too high to be included in Fig. 6.3. Their EERs are numerically recorded in Table 6.1. This experiment shows that FDA and DLDA suffer from severe over-fitting problem. Although ML performs better than in the first experiment, it still underperforms the traditional PCAE for some operating points. Fig. 6.3 shows again that the proposed MES method consistently achieves the lowest FAR and FRR among the 9 tested approaches for all different operating points.

Figure 6.3: False rejection rate against the false acceptance rate on the FERET database of 994 training images (497 subjects), 697 probe genuine images (697 subjects) and 697 probe imposter images (697 subjects). The total number of genuine matches is $697 \times 1 = 697$ and the total number of imposter matches is $697 \times 696 = 4,85,112$.

6.4.3 Results on FERET Database 2

This database is constructed, similar to one data set used in [110], by choosing 256 subjects with at least four images per subject. However, we use the same number of images (four) per subject for all subjects. 600 images of the first 200 subjects are used for training and also as gallery images. The remaining 200 images of 200 subjects serve as probe genuine images. Remaining 224 images are used as probe imposter images. The size of the normalized image is 150×130 , same as that in [110]. For such a large image size, we first apply PCA to remove the null space of $\mathbf{S}^t$ and then apply the MES approach on the 599-dimensional feature vectors. For this database we conducted 4 runs of training and testing with distinct probe genuine image set in each run. More specifically, the i^{th} images ($i = 1, 2, 3, 4$) of all training subjects are chosen to form probe genuine set and the remaining 3 images per subject serve as the training and gallery images. The total number of genuine matches is $200 \times 3 = 600$ and the total number of imposter matches is $4 \times 3 \times 56 \times 200 = 1,34,400$. Fig. 6.4 shows the ROC curves of the 4 runs of training and testing.

The ROC curves of FDA and ML do not appear in Fig. 6.4 because their FRRs and FARs are so high that their values are out of the range of Fig. 6.4. Their EERs are numerically recorded in Table 6.1. Although the second lowest ROC curve is different from the previous two experiments, Fig. 6.4 shows once more that the proposed MES method consistently delivers the most accurate face verification for all different operating points.

For an accurate record, various ERRs (in %) obtained from the above three experiments are numerically recorded in Table 6.1. It clearly demonstrates the superior performance of the proposed MES approach to all other approaches tested in the experiments on three different face databases.

Figure 6.4: False rejection rate against the false acceptance rate on the FERET database of 600 training images (200 subjects), 200 probe genuine images (200 subjects) and 224 probe imposter images (56 subjects). The total number of genuine matches is $200 \times 3 = 600$ and the total number of imposter matches is $56 \times 4 \times 200 \times 3 = 1,34,400$.

Table 6.1: Equal Error Rate (ERR %) of various approaches for three different databases.

Databases	PCAE	PCAM	FDA	NDA	DLDA	ML	UFS	DSL	MES
AR	34.2	21.9	13.2	5.0	5.2	19.1	4.3	5.9	2.6
FERET1	5.1	6.0	35.0	4.0	27.0	7.9	3.0	3.8	2.0
FERET2	13.0	15.5	34.8	3.6	10.5	23.8	2.7	2.5	2.0

6.5 Summary

We have conducted 3 sets of experiments on 3 different face databases that evaluate 9 subspace based approaches for the face verification task. Unlike face identification experiments where some sample images of all probe subjects can be included in the training, in all verification experiments of this work, the subjects of the probe imposters are excluded in the training. Moreover, in FERET database 1, the training subjects are different from those in the gallery and probe sets. The experimental results verify the difference in terms of accuracy between the face verification and the face identification. Methods that work well for the face identification may not necessarily do the same for the face verification task. ML is a good example for this. It is thus useful to test the verification performances of various approaches that were developed and tested for identification task.

The experiments show that UFS, NDA and DSL approaches perform better than PCAE, PCAM, FDA, DLDA and ML approaches. UFS keeps only a small principal subspace with largest eigenvalues for the discriminant evaluation. It suppresses more noise and thus has less over-fitting problem comparing to the FDA and DLDA that perform the discriminant evaluation in the whole range space. The good performance of NDA verifies that the null space contains important discriminative information and should not be simply discarded in the feature extraction. Another property of NDA is that it does not scale the features by the eigenvalues. This is one possible reason why NDA has better generalization than FDA, DLDA and ML. DSL extract two sets of features, one from a principal subspace and the other from its complementary subspace including the null space. Its relative good performance shows that the discriminative information resides in the both subspaces.

However, none of the three better approaches, UFS, NDA and DSL can consistently achieve the second best performance in the three experiments. One reason could be that all of them are suboptimal that extract features by the discriminant evaluation in a subspace or separately in two subspaces. The proposed MES

method shows superior verification performances to all the other 8 subspace based approaches. In all three experiments on the different face databases, the proposed MES method consistently achieves the lowest error rates at all different operating points. It is important to test a verification system at different operating points because there is no optimal threshold for a verification system and different applications in practice has different requirement of FAR and FRR. The superior verification performance of the proposed method is attributed to the eigenspectrum modeling that enables a global optimization by the discriminant evaluation in the whole space and alleviates the over-fitting problem by replacing the unreliable or noise sensitive small and zero eigenvalues by the modeled values. Therefore, the extracted features are the most discriminative in the whole space and stable or less sensitive to the noise disturbance, the data dimensionality and the number of training samples. Extensive experiments on three face databases demonstrate that the proposed approach consistently outperforms other 8 popular subspace based approaches tested in this chapter.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we have presented algorithms for an important and noteworthy application of computer vision and pattern recognition which is face recognition. We have proposed new approaches that are useful for improving the accuracy of the two aspects of face recognition - identification and verification. The main contributions of the thesis are to explore, analyze and propose solutions to the problems of feature extraction and regularization/scaling that exist in most of the computer vision problems such as face recognition. These proposed algorithms are experimentally evaluated and examined to be efficient, robust and can perform better feature scaling and extraction as compared to other methods. All these characteristics of our algorithms are useful for discrimination/classification purposes. Furthermore, discriminant evaluation of the eigenspace is an important aspect of classification methodologies, our proposed algorithms work on fresh ideas of whole space discriminant evaluations and feature extractions.

We began by considering two popular recent face recognition surveys [3, 24]. According to these surveys, PCA, FDA, NDA, ML and their variations are probably the best known subspace face recognition algorithms. We have investigated that

ML lacks optimal estimation of eigenvalues in the complementary portion of the eigenspectrum because the lower portion of the eigenspectrum (the variances) obtained from the limited training data are unreliable and their average causes severe over-fitting problem. We have proposed an ‘upper bound’ ρ selection mechanism, which attenuates the problems of unreliable small and zero eigenvalues caused by noise and the limited number of training samples. The superiority of our proposed approach over the existing ML method is demonstrated on large FERET database.

The new regularization and extraction strategy proposed in this thesis will probably help in further advancement of the general appearance based face recognition research. This is evident from the extensive experimentations and analysis on many large face image databases. The regularization procedure uses an eigen-model scheme and achieves more accuracy gains for smaller number of features and for smaller size of training set as compared to all other popular face recognition methods. This verifies that the proposed approach is more efficient in the feature extraction and more stable, less over-fitting the training data or better generalization. We have also studied the kernel subspace based face recognition. The new eigenratio scheme helps us to decompose the kernel eigenspectrum. The proposed weighting function circumvents the undue scaling of the eigenvectors corresponding to the unreliable small and zero eigenvalues and facilitates the feature extraction by the discriminant evaluation in the whole kernel eigenspace. Extensive experiments on popular face image databases demonstrate that our proposed approach consistently outperforms all other popular kernel based methods.

Finally, we have studied the intrinsic difference between face identification and verification. We have seen that a learning algorithm may suffer more severe over-fitting problem (or poor generalization) for face verification task than for the face identification and this leads to quite different evaluation results between face verification and identification systems. The proposed feature weighting scheme uses regularization of the partial eigenspectrum with the model eigenvalues and a constant model eigenvalue to replace the very small and zero eigenvalues. It experimentally

evaluates eight popular subspace based approaches for the face verification task and compares them with the proposed method. Experimental results on several popular face databases demonstrate that the proposed method consistently outperforms others.

7.2 Future Work

In this section, we present some research directions which can be pursued in the future. Based on our findings, all popular work (like PCA and FDA) on face recognition select the eigenvectors (or eigenfeatures) based on the variances of the sample covariance matrix. Future research should address issues related to what is the basis of selection of discriminative features not just based on the variances of the sample data. In face recognition problems, most of the methods use sample covariance matrices, which are always estimated from the limited number of training samples. Singularity of these covariance matrices are very common problems in pattern recognition. How to improve the generalization of these covariance matrices or overcome this singularity problem, which would help in classification is an interesting problem.

A common problem to all the holistic or appearance based face recognition methods as compared to the component based methods is that they are sensitive to large variations in pose and illumination conditions. How to handle large pose variations using appearance based face models is indeed a very challenging problem. Although some efforts are made to attenuate illumination effects using some pre-processing of face images their applications along with large pose variations are still challenging and are largely unsolved in the present literature.

The present studies show that holistic or appearance based approaches have dominated the face recognition systems and have shown better performance than component based approaches. In general, these holistic or appearance based approaches use variance information from sample training samples. These 2D vari-

ance information may not be much useful in handling large pose changes fueled by self (or other) occlusions and in combination with extreme illumination changes. These problems can be alleviated by using local features methods or sophisticated face modeling schemes. One open question is how to arbitrate the use of holistic and local features or face modeling. Since the later procedures are dependent on many parameter selections and face/camera measurements. How to determine and use the features (local/global) remains an open problem. One suggestion is to explore different feature spaces and find some discriminative features using some disparity criteria. In particular, one of the most interesting questions is how “local” or “global” features should be to get the best recognition performance.

Based on our findings it is evident that the kernel mapping of the input data to some nonlinear plane may or may not be helpful in further classification. The area that merits further investigation is how to choose the optimal mapping function and its various parameters so that the mapping facilitates in face classification. These are very challenging in present machine learning literature and are under investigation. Modeling face variation under realistic settings is still challenging and are largely unsolved in the literature. For example, outdoor environments, natural aging, etc.

The results we obtained after regularization of the eigenspectrum are promising. Exploring the whole eigenspace can improve classification abilities. However, how to treat the null space which is believed to contain most discriminative information is ambiguous because we don't have any knowledge about its real variances. There are elements in our algorithms that make them applicable for these face identification and verification tasks. In particular, the fact that we try to model the variances of the whole eigenspace rather than using the real variances is probably a better approach.

Bibliography

- [1] L. Wiskott, J. Fellous, and V. Malsburg, “Face recognition by elastic bunch graph matching,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, July 1997.
- [2] B. Moghaddam, T. Jebara, and A. Pentland, “Bayesian face recognition,” *Pattern Recognition*, vol. 33, no. 11, pp. 1771–1782, November 2000.
- [3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, December 2003.
- [4] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, second edition, 1991.
- [5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. New York: John Wiley and Sons, 2001.
- [6] G. Shakhnarovich and B. Moghaddam, *Face recognition in subspaces*. 201 Broadway, Cambridge, Massachusetts 02139: Springer, March 2005.
- [7] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of the face recognition grand challenge,” in *IEEE Conf. Comp. Vis. & Patt. Recog.*, San Diego, June 2005, pp. 947–954.
- [8] P. J. Phillips, W. Scruggs, A. O’Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe, “The Face Recog-

- nition Vendor Test (FRVT) 2006,” *FRVT 2006 Report*:
<http://face.nist.gov/frvt/frvt2006/FRVT2006andICE2006LargeScaleReport.pdf>,
 2006.
- [9] A. Abate, M. Nappi, D. Riccio, and G. Sabatino, “2D and 3D face recognition: A survey,” *Pattern Recognition Letters*, vol. 28, pp. 1885–1906, January 2007.
 - [10] K. Bowyer, K. Chang, and P. Flynn, “A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition,” *Computer Vision and Image Understanding*, vol. 101, pp. 1–15, 2006.
 - [11] J. Yang, D. Zhang, and J. Y. Yang, “Globally Maximizing, Locally Minimizing: Unsupervised Discriminant Projection with Applications to Face and Palm Biometrics,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 650–664, April 2007.
 - [12] W. Liu, Y. Wang, S. Z. Li, and T. N. Tan, “Null space approach of Fisher discriminant analysis for face recognition,” in *ECCV workshop on Biometric Authentication*, Prague, Czech Republic, May 2004, pp. 32–44.
 - [13] X. Wang and X. Tang, “A unified framework for subspace face recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1222–1228, September 2004.
 - [14] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, “Discriminative common vectors for face recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4–13, January 2005.
 - [15] T. K. Kim and J. Kittler, “Locally linear discriminant analysis for multimodally distributed classes for face recognition with single modal image,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 318–327, March 2005.

- [16] C. Liu and H. Wechsler, “Enhanced Fisher linear discriminant models for face recognition,” in *Int. Conf. on Pattern Recognition (ICPR)*, Queensland, Australia, August 1998, pp. 1368–1372.
- [17] Q. S. Liu, X. Tang, H. Lu, and S. D. Ma, “Face recognition using kernel scatter-difference-based discriminant analysis,” *IEEE Trans. Neural Network*, vol. 17, no. 4, pp. 1081–1085, July 2006.
- [18] Q. Yang and X. Tang, “Recent advances in subspace analysis for face recognition,” in *Sinobiometrics, LNCS Vol. 3338*, Berlin, 2004, pp. 275–287.
- [19] S. Fidler, D. Skocaj, and A. Leonardis, “Combining Reconstructive and Discriminative Subspace Methods for Robust Classification and Regression by Subsampling,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 337–350, March 2006.
- [20] M. Zhu and A. Martinez, “Subclass discriminant analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1274–1286, August 2006.
- [21] D. Voth, “In the news: Face recognition technology,” *IEEE magazine on intelligent systems*, vol. 18, no. 3, pp. 4–7, May-June 2003.
- [22] K. Kendrick, A. Costa, A. Leigh, M. Hinton, and J. Peirce, “Sheep don’t forget a face,” *Nature*, vol. 411, pp. 165–166, 2001.
- [23] M. A. Peterson and G. Rhodes, *Perception of faces, objects and scenes*. 198 Madison Avenue, New York-10016, USA: Oxford University Press, 2003.
- [24] R. Chellappa, C. L. Wilson, and S. Sirohey, “Human and machine recognition of faces: A survey,” *Proceedings of IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.
- [25] L. Sirovich and M. Kirby, “Low-dimensional Procedure for the Characterization of Human Faces,” *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, vol. 4, no. 3, pp. 519–524, March 1987.

- [26] M. Kirby and L. Sirovich, "Application of Karhunen-Loeve procedure for the characterization of Human faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, January 1990.
- [27] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, January 1991.
- [28] B. Scholkopf, A. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
- [29] B. Scholkopf, S. Mika, C. J. Burges, P. Knirsch, K. R. Muller, G. Ratsch, and a. Smola, "Input space versus feature space in kernel-based methods," *IEEE Trans. Neural Network*, vol. 10, pp. 1000–1017, September 1999.
- [30] R. A. Fisher, "The use of Multiple Measures in Taxonomic Problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.
- [31] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, August 1996.
- [32] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, July 1997.
- [33] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, February 2001.
- [34] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," *Pattern Recognition*, vol. 34, no. 10, pp. 2067–2070, October 2001.

- [35] D. Q. Dai and P. C. Yuen, "Regularized discriminant analysis and its application to face recognition," *Pattern Recognition*, vol. 36, no. 3, pp. 845–847, March 2003.
- [36] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small size problem of LDA," *Proc. Int. Conf. on Pattern Recognition (16th)*, vol. 3, pp. 29–32, August 2002.
- [37] J. Yang, D. Zhang, and J. Y. Yang, "A generalized K-L expansion method which can deal with small sample size and high-dimensional problems," *Pattern Analysis & Applications*, vol. 6, no. 1, pp. 47–54, April 2003.
- [38] L. F. Chen, H. Y. M. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, October 2000.
- [39] Y. Q. Cheng, Y. M. Zhuang, and J. Y. Yang, "Optical Fisher discriminant analysis using the rank decomposition," *Pattern Recognition*, vol. 25, pp. 101–111, 1992.
- [40] Z. Q. Hong and J. Y. Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane," *Pattern Recognition*, vol. 24, pp. 317–324, 1991.
- [41] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 623–627, June 2000.
- [42] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Muller, "Fisher discriminant analysis with kernels," *IEEE Workshop on Neural Networks for Signal Processing IX*, pp. 41–48, 1999.
- [43] C. Liu and H. Wechsler, "Evolutionary pursuit and its application to face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 570–582, 2000.

- [44] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, July 1997.
- [45] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, October 2000.
- [46] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.
- [47] M. Bartlett, J. Movellan, and T. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Network*, vol. 13, no. 6, pp. 1450–1464, November 2002.
- [48] K. Kwak and W. Pedrycz, "Face recognition using an enhanced independent component analysis approach," *IEEE Trans. Neural Network*, vol. 18, no. 2, pp. 530–541, March 2007.
- [49] B. Draper, K. Baek, M. Bartlett, and J. Beveridge, "Recognizing faces with PCA and ICA," *Computer Vision and Image Understanding*, vol. 91, pp. 115–137, 2003.
- [50] H. Ekenel and B. Sankur, "Feature selection in the independent component subspace for face recognition," *Pattern Recognition Letters*, vol. 25, pp. 1377–1388, 2004.
- [51] T. Sintra, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, April 1994.
- [52] P. Yuen and J. Lai, "Face representation using independent component analysis," *Pattern Recognition*, vol. 35, no. 6, pp. 1247–1257, June 2002.
- [53] S. Laughlin, "A simple coding procedure enhances a neuron's information capacity," *Z. Naturforsch*, vol. 36, pp. 910–912, 1981.

- [54] A. Bell and T. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computing*, vol. 7, no. 6, pp. 1129–1159, 1995.
- [55] J. Nadal and N. Parga, “Non-linear neurons in the low noise limit: A factorial code maximizes information transfer,” *Network*, vol. 5, pp. 565–581, 1994.
- [56] F. Bach, “Kernel independent component analysis,” *The Journal of Machine Learning Research*, vol. 3, pp. 1–48, March 2003.
- [57] A. Bell and T. Sejnowski, “The independent components of natural scenes are edge filters,” *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.
- [58] K. Baek, B. Draper, J. Beveridge, and K. She, “PCA vs ICA: A comparison on the FERET data set,” in *International Joint Conference on Information Sciences*, Durham, North Carolina, USA, March 2002, pp. 824–827.
- [59] B. Moghaddam, “Principal manifolds and Bayesian subspaces for visual recognition,” in *Inter. Conf. on Computer Vision (ICCV’99)*, September 1999, pp. 1131–1136.
- [60] M. Vasilescu and D. Terzopoulos, “Multilinear analysis of image ensembles: Tensorfaces,” in *In the Proceedings of European Conference on Computer Vision*, CopenhagenDenmark, May 2002, pp. 447–460.
- [61] ———, “Multilinear subspace analysis of image ensembles,” in *In the Proceedings of IEEE Computer Vision and Pattern Recognition*, Madison, WI, June 2003, pp. 93–99.
- [62] S. Lin, S. Kung, and L. Lin, “Face recognition/detection by probabilistic decision-based neural network,” *IEEE Trans. Neural Network*, vol. 8, pp. 114–132, 1997.
- [63] S. Kung and J. Taur, “Decision-based neural networks with signal/imageclassification applications,” *IEEE Trans. Neural Network*, vol. 6, pp. 170–181, 1995.

- [64] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.
- [65] P. J. Phillips, M. Kearns, S. Solla, and D. Cohn, "Support vector machines applied to face recognition," in *Advances in Neural Information Processing Systems*, vol. 11, 1998.
- [66] K. Jonsson, J. Matas, J. Kittler, and Y. Li, "Learning support vectors for face verification and recognition," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 208–213.
- [67] G. Guo, S. Z. Li, and K. Chan, "Face recognition by support vector machines," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 96–201.
- [68] G. Cui and W. Gao, "Support vector machines for face recognition with two-layer generated virtual data," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, vol. 3, Aug 2004, pp. 570–573.
- [69] L. Wiskott and C. Malsburg, "Face recognition by dynamic link matching," <http://nn.cs.utexas.edu/web-pubs/htmlbook96/wiskott/>, 1996.
- [70] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. Malsburg, R. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Computers*, vol. 42, no. 3, pp. 300–311, March 1993.
- [71] J. Buhmann, M. Lades, and C. Malsburg, "Size and distortion invariant object recognition by hierarchical graph matching," in *International Joint Conference on Neural Networks*, vol. 2, June 1990, pp. 411–416.
- [72] D. Bolme, "MS thesis on EBGM, CSU, USA," *Technical report*: http://www.cs.colostate.edu/evalfacerec/papers/EBGMThesis_Final.pdf, June, 2003.

- [73] R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, “The CSU Face Identification Evaluation System Users Guide: Version 5.0,” *Technical Report: <http://www.cs.colostate.edu/evalfacerec/data/normalization.html>*, 2003.
- [74] I. Biederman and P. Kalocsai, *Face recognition: From theory to applications*. Berlin, Germany: Springer-Verlag, 1998.
- [75] R. Brunelli and T. Poggio, “Face recognition: Features versus templates,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042–1052, 1993.
- [76] A. Nefian and M. Hayes, “An embedded HMM-based approach for face detection and recognition,” in *Proceedings of IEEE Conf. on Acoustic, Speech and Signal Processing*, vol. 6, 1999, pp. 3553–3556.
- [77] A. M. Martinez, “Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, June 2002.
- [78] A. Pentland, B. Moghaddam, and T. Starner, “View-based and modular eigenspaces for face recognition,” *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1994.
- [79] P. Penev and J. Atick, “Local feature analysis: A general statistical theory for object representation,” *Network: Computation in Neural Systems*, vol. 7, pp. 477–500, 1996.
- [80] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, “Active shape models—their training and application,” *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, January 1995.
- [81] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” in *First Int. Conf. on Computer Vision*, 1987, pp. 259–268.

- [82] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.
- [83] Z. Xue, S. Z. Li, D. Shen, and E. Tcho, "A novel Bayesian shape model for facial feature extraction," in *Seventh International Conference on Control, Automation, Robotics And Vision*, Singapore, Dec 2002, pp. 514–519.
- [84] S. Z. Li and J. Lu, "Modelling Bayesian estimation for deformable contours," in *7th IEEE Conf. on Computer Vision*, Greece, May 1999, pp. 991–996.
- [85] H. Ip and D. Shen, "An affine-invariant active contour model (AI-snake) for model-based segmentation," *Image and Vision Computing*, vol. 16, no. 2, pp. 125–146, 1998.
- [86] M. Teixeira, "The Bayesian intrapersonal/extrapersonal classifier," *CSU MS Thesis*: <http://www.cs.colostate.edu/evalfacerec/papers/teixeiraThesis.pdf>, July, 2003.
- [87] B. Moghaddam, "Principal manifolds and probabilistic subspace for visual recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 780–788, June 2002.
- [88] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, Washington DC, USA, June 2004, pp. 564–569.
- [89] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 195–200, January 2003.
- [90] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, February 2003.

- [91] T. K. Kim, H. Kim, W. Hwang, and J. Kittler, "Independent component analysis in a local facial residue space for face recognition," *Pattern Recognition*, vol. 37, pp. 1873–1885, September 2004.
- [92] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1977–1981, December 2005.
- [93] Q. Liu, H. Lu, and S. Ma, "Improving kernel Fisher discriminant analysis for face recognition," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 42–49, January 2004.
- [94] W. S. Chen, P. C. Yuen, J. Huang, and D. Q. Dai, "Kernel machine-based one-parameter regularized fisher discriminant method for face recognition," *IEEE Trans. Systems, Man, and Cybernetics-Part B*, vol. 35, no. 4, pp. 659–669, August 2005.
- [95] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, February 2005.
- [96] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant analysis for face recognition," *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 467–476, April 2002.
- [97] C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 572–581, May 2004.
- [98] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, March 2005.

- [99] H. Gao and J. W. Davis, "Why direct LDA is not equivalent to LDA," *Pattern Recognition*, vol. 39, pp. 1002–1006, 2006.
- [100] K. Liu, Y. Q. Cheng, J. Y. Yang, and X. Liu, "An efficient algorithm for Foley-Sammon optical set of discriminant vectors by algebraic method," *International Journal of Pattern Recognition Artificial Intelligence*, vol. 6, pp. 817–829, 1992.
- [101] J. Yang and J. Y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognition*, vol. 36, pp. 563–566, 2003.
- [102] X. Zhuang and D. Dai, "Improved discriminant analysis for high-dimensional data and its application to face recognition," *Pattern Recognition*, vol. 40, pp. 1570–1578, 2007.
- [103] C. Liu and H. Wechsler, "Robust coding schemes for indexing and retrieval from large face database," *IEEE Trans. Image Processing*, vol. 9, no. 1, pp. 132–137, January 2000.
- [104] W. Liu, Y. W. Wang, S. Z. Li, and T. N. Tan, "Null space-based kernel Fisher discriminant analysis for face recognition," *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 369–374, 2004.
- [105] B. Zhang, H. Zhang, and S. Ge, "Face recognition by applying wavelet sub-band representation and kernel associative memory," *IEEE Trans. Neural Network*, vol. 15, no. 1, pp. 166–177, 2004.
- [106] V. Perlibakas, "Face recognition using principal component analysis and wavelet packet decomposition," *Informatica*, vol. 15, no. 2, pp. 243–250, 2004.
- [107] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, March 1989.
- [108] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized discriminant analysis for the small sample size problem in face recognition," *Pattern Recognition Letters*, vol. 24, pp. 3079–3087, 2003.

- [109] ———, “Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition,” *Pattern Recognition Letters*, vol. 26, pp. 181–191, 2005.
- [110] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, and S. Z. Li, “Ensemble-based discriminant learning with boosting for face recognition,” *IEEE Trans. Neural Networks*, vol. 17, no. 1, pp. 166–178, January 2006.
- [111] F. Samaria and A. Harter, “Parameterization of a stochastic model for human face identification,” in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, USA, December 1994, pp. 138–142.
- [112] B. G. Park, K. M. Lee, and S. U. Lee, “Face recognition using face-ARG matching,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1982–1988, December 2005.
- [113] Georgia tech face database. [Online]. Available: http://www.anefian.com/face_reco.htm.
- [114] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. J. Smola, K. R. Muller, S. A. Solla, and T. K. Leen, “Invariant feature extraction and classification in kernel spaces,” *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, vol. 12, pp. 526–532, 2000.
- [115] H. Gupta, A. K. Agarwal, T. Pruti, C. Shekhar, and R. Chellappa, “An experiment evaluation of linear and kernel-based methods for face recognition,” *IEEE Workshop on Applications on Computer Vision*, December 2002.
- [116] M. H. Yang, “Kernel Eigenfaces vs. kernel Fisherfaces: Face recognition using kernel methods,” *Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 215–220, May 2002.
- [117] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Face recognition using kernel direct discriminant analysis algorithms,” *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 117–126, January 2003.

- [118] M. Kirby, *Dimensionality reduction and pattern analysis: An empirical approach*. Wiley, 2000.
- [119] Q. S. Liu, R. Huang, H. Q. Lu, and S. D. Ma, “Face Recognition Using Kernel based Fisher Discriminant Analysis,” *Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 197–201, May 2002.
- [120] N. A. M Yang and D. Kriegman, “Face recognition using kernel eigenfaces,” in *IEEE Conf. on Image Processing*, 2000, pp. 37–40.
- [121] K. J. K. Kin and H. Kim, “Face recognition using kernel principal component analysis,” *IEEE Signal Processing Letters*, vol. 9, pp. 40–42, Feb 2002.
- [122] S. Lucey and T. Chen, “A GMM parts based face representation for improved verification through relevance adaptation,” in *Proc. Computer Vision and Pattern Recognition*, 2004, pp. 855–861.
- [123] H. Liu, C. Su, Y. Chiang, and Y. Hung, “Personalized face verification system using owner-specific cluster-dependent LDA-subspace,” in *Int. Conf. on Pattern Recognition*, 2004, pp. 344–347.
- [124] M. Sadeghi and J. Kittler, “Decision making in the LDA space: Generalised gradient direction metric,” in *Proceedings of 6th IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2004, pp. 248–253.
- [125] A. Bazin and M. Nixon, “Facial verification using probabilistic methods,” in *British Machine Vision Association Workshop on Biometrics*, London, August 2004.

Publications

- (1) B. Mandal, X. D. Jiang and A. Kot, "Face Verification Using Modeled Eigenspectrum," The Open Artificial Intelligence Journal, Bentham Open, DOI: 10.2174/1874061800802010035, (Accepted for publication), 19 May 2008.
- (2) X. D. Jiang, B. Mandal and A. Kot, "Complete Discriminant Evaluation and Feature Extraction in Kernel Space for Face Recognition," Machine Vision and Applications (MVA), Springer-Verlag, DOI: 10.1007/s00138-007-0103-1, full length regular paper accepted for publication 21st July, 2007.
- (3) X. D. Jiang, B. Mandal and A. Kot, "Eigenfeature Regularization and Extraction in Face Recognition," IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI), vol. 30, no. 3, pp. 383-394, March 2008.
- (4) X. D. Jiang, B. Mandal and A. Kot, "Enhanced Maximum Likelihood Face Recognition," IEE Electronics Letters, vol. 42, no. 19, pp. 1089-1090, September 2006.
- (5) B. Mandal, X. D. Jiang and A. Kot, "Verification of Human Faces Using Predicted Eigenvalues," International Conference on Pattern Recognition (ICPR 2008), Tempa, Florida, USA, (Accepted for oral presentation) Dec 2008.
- (6) B. Mandal, X.D. Jiang and A. Kot, "Dimensionality Reduction in Subspace Face Recognition," IEEE Sixth International Conference on Information, Communications and Signal Processing (ICICS 2007), Singapore, pp. 1-5, 10-13 December 2007.

- (7) B. Mandal, X. D. Jiang and A. Kot, "Kernel Fisher Discriminant Analysis in Full Eigenspace," International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV 2007), Las Vegas, Nevada, USA, pp. 235-241, 25-28 June 2007.
- (8) X.D. Jiang, B. Mandal and A. Kot, "Face Recognition Based on Discriminant Evaluation in the Whole Space," IEEE 32nd International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Honolulu, Hawaii, USA, pp. 245-248, 15-20 April 2007.
- (9) B. Mandal, X.D. Jiang and A. Kot, "Multi-scale Feature Extraction for Face Recognition," IEEE International Conference on Industrial Electronics and Applications (ICIEA 2006), Singapore, pp. 1-6, 24-26 May 2006 (invited paper).
- (10) X.D. Jiang, B. Mandal and A. Kot, "Improved Bayesian Approach for Face Recognition," IEEE Fifth International Conference on Information, Communications and Signal Processing (ICICS 2005), Bangkok, Thailand, pp. 162-166, 6-9 December 2005.