Subspace analysis for face recognition

Kong, Hui

2006

Kong, H. (2006). Subspace analysis for face recognition. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/39102>

<https://doi.org/10.32657/10356/39102>

Subspace Analysis for Face Recognition

Kong Hui

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2006

Statement of Originality

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

DEC-29-2006
.....
Date

.....
Kong Hui

Acknowledgments

I would like to express my deepest appreciation to my supervisor, Associate Professor Teoh Eam Khwang, and co-supervisor, Dr Ronda Venkateswarlu, for their enthusiastic, patient guidance, encouragement and many invaluable comments and suggestions throughout the entire PhD research project. Without their assistance and direction, this research would not be possible.

I am also sincerely and deeply grateful to the following persons: (1) Professor Matthew Turk of Department of Computer Science, University of California, Santa Barbara, USA; (2) Dr Jian-Gang Wang from Institute for Infocomm Research, Singapore; (3) Dr Lei Wang from Research School of Information, Science and Engineering, Australian National University, Australia; and (4) Mr Xuchun Li from Intelligent Machine Research Lab (*IMRL*) of School of Electrical and Electronic Engineering (*EEE*), Nanyang Technological University (*NTU*), Singapore.

Special thanks also go to all the staff members of *IMRL*, School of *EEE*, *NTU*, for their generous assistance throughout the entire project.

I wish to acknowledge my family members who showed so much concern and care for me in the course of my studies. Their encouragement and unyielding support have contributed to the completion of this thesis. Heartfelt thanks to my wife, Ms. Ling Ding, a future Financial Analyst and Economist from University of California, Santa Cruz, USA. This thesis is in memory of my mother, Chunying Lv, the greatest woman in my life. Finally, thanks to the Nanyang Technological University, Singapore, for providing the research scholarship.

Summary

This thesis presents a research project on face recognition via subspace analysis algorithms. Although face recognition has been actively studied over the past decade, the state-of-the-art recognition systems cannot yield satisfying performance due to small number of training samples available and image variations caused illumination, pose and others. In this thesis, we propose novel subspace analysis approaches that focus on how to recognize human faces with only few training samples.

Generally, each image is represented as a column vector by concatenating all the rows of the image. Therefore, the resulting image vector lies in a high dimensional space. However, the high dimensionality will induce a lot of problems such as the high storage-space requirement and computational complexity. Most important of all, it will degrade the recognition accuracy because of the Curse of Dimensionality. On the other hand, it has been reported that the face images can be represented in a low dimensional subspace of face space. Therefore, the dimension reduction is a crucial step no matter for the efficiency or the accuracy of the recognition algorithms. Conventional dimension reduction schemes such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) seek the low dimensional representations of the face images by individual subspace analysis, and these extracted low dimensional feature vectors can be used for recognition purposes.

However, the performance of these conventional subspace analysis approaches is limited by the number of training samples available. Especially when there only exist few training samples for each subject, the Small Sample Size (SSS) problem

is inevitable for LDA method. To deal with the SSS problem, three discriminant subspace analysis methods are given in the second, third and fourth contributions of this thesis.

In the first contribution, inspired by the recent 2D Principal Component Analysis (2DPCA), a framework of Generalized 2D Principal Component Analysis (G2DPCA) is proposed. In the second contribution, a framework of 2D Fisher Discriminant Analysis (2DFDA) is proposed. This framework contains Unilateral 2DFDA (U2DFDA), Bilateral 2DFDA (B2DFDA) and Kernel-based 2DFDA (K2DFDA). In the third contribution, a framework of Discriminant Low dimensional Subspace Analysis (DLSA) method is proposed for face recognition with small number of training samples. Two algorithms, namely Unified Linear Discriminant Analysis (ULDA) and Modified Linear Discriminant Analysis (MLDA), are proposed in this framework. In the fourth contribution, by analyzing the problems in the current FLD-based approaches, a framework of Ensemble Learning for Diversified Fisher Linear Discriminant (ELDFLD) is proposed.

In summary, this thesis presents four novel algorithms for face recognition using subspace analysis. All these algorithms can be applied to the situation where there are limited number of training samples. The second, third and fourth algorithms are also tolerable to variations caused by illumination, pose, expression etc., because they all based on discriminant analysis.

Contents	v

Table of Contents

Acknowledgements	i
Summary	ii
List of Figures	xvi
List of Tables	xviii
List of Symbols	xix
1 Introduction	1
1.1 Motivation	2
1.2 Objectives	5
1.3 Original Contributions	6
1.4 Thesis Organization	8
2 Face Recognition – A Literature Review	11
2.1 Introduction	11
2.2 Face Recognition Based on Subspace Analysis	12

Contents	vi
2.2.1 Principal Component Analysis	13
2.2.2 Fisher Linear Discriminant	14
2.2.3 Kernel Principal Component Analysis	16
2.2.4 Kernel Fisher Discriminant Analysis	17
2.3 Face Recognition from Still Images	19
2.3.1 Holistic Approaches	20
2.3.2 Feature-Based Approaches	25
2.3.3 Hybrid Approaches	27
2.4 Face Recognition from Video Sequences	29
2.5 Pose- and Illumination-invariant Face Recognition Algorithms	35
2.5.1 Pose-invariant Face Recognition Algorithms	35
2.5.2 Illumination-invariant Face Recognition Algorithms	40
2.6 A Summary of Face Databases Used in this Thesis	45
2.7 Concluding Remarks	45
3 Generalized 2D Principal Component Analysis	47
3.1 Introduction	47
3.2 2D Principal Component Analysis	49
3.3 The Essence of 2D Principal Component Analysis	50
3.4 Bilateral 2D Principal Component Analysis	52
3.4.1 The Algorithm	52
3.4.2 Image Representation Using B2DPCA	56

Contents	vii
3.4.3 Face Recognition Using B2DPCA	57
3.5 Kernel 2D Principal Component Analysis	57
3.6 Experimental Results and Discussion	61
3.6.1 Face Recognition on ORL and UMIST Face Databases	61
3.6.2 Effect of d -value on Recognition Rate	64
3.6.3 Convergence Analysis	65
3.6.4 Reconstruction and Compression of Face Images	66
3.7 Concluding Remarks	67
4 2D Fisher Discriminant Analysis	69
4.1 Introduction	69
4.1.1 The Small Sample Size Problem	69
4.1.2 Previous Solutions to the SSS problem	70
4.1.3 2D Data Representation Model	72
4.1.4 2D Fisher Discriminant Analysis	73
4.2 Unilateral 2D Fisher Discriminant Analysis	74
4.3 Bilateral 2D Fisher Discriminant Analysis	77
4.4 The Essence of 2D Fisher Discriminant Analysis	79
4.5 Kernel 2D Fisher Discriminant Analysis	80
4.6 Experimental Results and Discussion	83
4.6.1 Experiments on ORL and YaleB face databases	85
4.6.2 Performance Comparison on UMIST Database	95

Contents	viii
4.7 Concluding Remarks	96
5 Discriminant Low-dimensional Subspace Analysis	99
5.1 Introduction	99
5.2 Unified Discriminant Subspace Analysis	101
5.3 Modified Linear Discriminant Analysis	106
5.4 Kernel Modified Fisher Discriminant Analysis	107
5.5 Experimental Results and Discussion	110
5.5.1 Random Grouping of Training and Testing Sets	112
5.5.2 Comparison of Recognition Performance	113
5.5.3 The Effect of α and β on Recognition rate	114
5.5.4 The Effect of μ on Recognition Rate	114
5.6 Concluding Remarks	115
6 Ensemble Learning for Diversified Fisher Linear Discriminant	119
6.1 Introduction	119
6.2 Problems in Single/Multiple-classifier System	120
6.2.1 Problems in Single-classifier System	120
6.2.2 Problems in Multi-classifier System	122
6.2.3 Our Solutions to the Problems in Single-classifier System . . .	123
6.2.4 Our Solutions to the Problems in Multi-classifier System . . .	124
6.3 Ensemble Learning for Diversified Fisher Linear Discriminant	125
6.3.1 Random Sub-feature LDA	125

Contents	ix
6.3.2 AdaBoost-LDA	126
6.3.3 Coupling of AdaBoost-LDA and RS-LDA for Diversified FLD	127
6.4 Experiment Results and Discussion	129
6.4.1 Performance of EL-DFLD with/without D-A Consideration .	130
6.4.2 Performance of AdaBoost-LDA	131
6.4.3 Performance by Coupling RS-LDA and AdaBoost-LDA	132
6.4.4 Performance Comparison with Previous Methods	132
6.5 A comparative study of the four proposed linear methods	133
6.5.1 Comparison of recognition performance on ORL and YaleB face databases	134
6.5.2 Comparison of computational requirement	135
6.6 Concluding Remarks	135
7 Conclusion and Recommendations	137
7.1 Conclusion	137
7.2 Recommendations for Future Research	139
Author's Publications	143
Bibliography	146

List of Figures

1.1	Illustration of the process of face recognition (face images adopted from CMU PIE [156] and YaleB [158] face image databases).	2
1.2	An illustration of the imaging system.	3
1.3	Image samples under different imaging conditions (face images are adopted from YaleB [158], ORL [155] and CMU AMP Expression [160] databases).	4
1.4	The face images of the same person under different illumination and pose conditions. (face images are adopted from CMU PIE [156] face databases).	5
2.1	A comparison of PCA and FLD for a two class problem where data for each class lies near a linear subspace (adopted from [30]).	15
2.2	Electronically modified images which were correctly identified (adopted from [34]).	20
2.3	Reconstructed images using 300 PCA projection coefficients for electronically modified images as shown in Figure 2.2 (adopted from [34]).	21
2.4	Comparison of dual- and standard-eigenface. (a) intrapersonal, (b) extrapersonal and (c) standard (adopted from [76]).	22

List of Figures	xii
2.5 Different projection bases. First row shows the first five pure LDA basis images. Second row shows the first five subspace LDA basis images. The average face and first four eigenfaces Φ are shown on the third row (adopted from [33]).	23
2.6 Two architectures for ICA. Left: architecture for finding statistically independent basis images. Right: architecture for finding a factorial code (adopted from [103]).	25
2.7 The bunch graph representation of faces used in elastic graph matching (adopted from [90]).	26
2.8 Comparison of matching: (a) test views, (b) eigenface matches, (c) eigenfeature matches (adopted from [14]).	28
2.9 The face recognition scheme based on flexible appearance model (adopted from [97]).	29
2.10 Active appearance model: Top row shows the appearances by varying the most significant identity parameters and bottom row shows the appearances by manipulating residual variation without affecting identity (adopted from [119]).	32
2.11 Identity surface (adopted from [115]).	34
2.12 The process of constructing the illumination cone. First row: The seven training images. Second row: Images corresponding to the columns of $\overline{\mathbf{B}}$. Third row: Reconstruction up to a GBR transformation. Fourth row: Synthesized images from the illumination cone of the face under novel lighting conditions but fixed pose (adopted from [70]).	36
2.13 Synthesized images under variable pose and lighting generated from the training images (adopted from [70]).	37

List of Figures	xiii
2.14 The best fit to a profile model is projected to the frontal model to predict new views (adopted from [136]).	39
2.15 View synthesis by parallel deformation. (A): the prototype flow is measured between the prototype image and the novel image at the same pose. (B): the flow is mapped onto the novel face. (C): the novel face is 2D-warped to the virtual view (adopted from [135]). . . .	40
2.16 In each row, the appearance of the same face is different under different lighting conditions (adopted from [157]).	41
2.17 Testing the invariance of the Quotient-image (Q-image) to varying illumination. Top row: Original images of a novel face taken under five different illuminations. Bottom row: The Q-images corresponding to the novel images, computed with respect to the bootstrap set of ten objects (adopted from [128]).	42
2.18 Image synthesis example. (a) shows the original image and (b) shows the quotient image of (a) from the $N = 10$ bootstrap set. (c) to (e) are the bootstrap set. (f) to (k) are the synthetic images (adopted from [128]).	43
2.19 The comparison of image rendering results. The original images are shown in the first column. The second column shows prototype images rendered using [131]. Prototype images rendered using symmetric SFS are shown in the third column. The fourth column shows real images that are close to the prototype images (adopted from [35]). . .	44
2.20 The first nine harmonic images of a face object (from left to right, top to bottom) (adopted from [132]).	45
3.1 Ten sample images of two subjects in ORL database	61

List of Figures	xiv
3.2 Eighteen sample images of subject 1a from UMIST face database labelled by #1, #2,..., #18 from left to right	61
3.3 The effect of different d -value on recognition rate of B2DPCA	64
3.4 Convergency of B2DPCA	65
3.5 First row: raw images. Second row and fourth row: image recon- structed and compressed by 2DPCA using 2 and 8 <i>principal com- ponent (vectors)</i> respectively. Third row and fifth row: image re- constructed and compressed by B2DPCA with $d = 10$ and $d = 20$ respectively.	66
4.1 Twenty sample images of two subjects in Yale face database B	84
4.2 Comparison between U2DFDA and B2DFDA with different number of <i>Fisher feature vectors</i> . (a)-(c): Two, three and four training sam- ples respectively for each subject on ORL database; (d)-(f): Two, three and four training samples respectively for each subject on Yale face database B.	87
4.3 Optimal number of <i>Fisher feature vector</i> and <i>Eigen feature vector</i> in U2DFDA and 2DPCA with different number of training samples for each subject on the ORL database	89
4.4 Optimal number of <i>Fisher feature vector</i> and <i>Eigen feature vector</i> in U2DFDA and 2DPCA with different number of training samples for each subject on the Yale face database B	89
4.5 Recognition rate of U2DFDA and B2DFDA compared with the other subspace methods on the ORL face database	90
4.6 Recognition rate of U2DFDA and B2DFDA compared with the other subspace methods on the Yale face database B	91

4.7	Performance of Fisherface with different dimension of the PCA subspace when the number of training sample is 4 for each subject in the ORL database	92
4.8	Optimal performance of FDA compared with optimal performance of U2DFDA and B2DFDA on the ORL database	93
4.9	Optimal performance of FDA compared with optimal performance of U2DFDA and B2DFDA on the Yale face database B	94
4.10	The comparison of performance of B2DFDA and U2DFDA with Ye's 2DLDA on ORL face database	94
4.11	The comparison of performance of B2DFDA and U2DFDA with Ye's 2DLDA on Yale face database B	95
5.1	Sample images in the combined database	111
5.2	Experimental results on the combined database	112
5.3	The effect of α and β on recognition rate	116
5.4	The performance of MLDA in recognition with different μ value. (a)-(c): Coarse adjustment of μ with two, three and four training samples respectively for each subject; (d)-(f): Fine adjustment of μ with two, three and four training samples respectively for each subject.	117
6.1	Recognition rate of Fisherface classifier with different dimension of PCA subspace.	122
6.2	The effect of our algorithm in dealing with D-A. (a): Comparison on ORL database; (b): Comparison on YaleB.	131
6.3	The comparison of performance of EL-DFLD with the other LDA-based methods in recognition. (a): Comparison on ORL database; (b): Comparison on YaleB.	133

List of Tables

2.1	Typical Applications of Face Recognition	12
2.2	A Summary of Face Databases Used in This Thesis	46
3.1	The algorithm for computing U_{Opt} and V_{Opt}	56
3.2	Experiments on ORL database: the first row indicates the number of training sample for each person and the first column indicates the different methods used.	62
3.3	Experiments on UMIST database: the first row indicates the training images for each person and the first column indicates the different methods used.	63
4.1	Recognition using 2DFDA	79
4.2	Five Experiments on ORL and YaleB	85
4.3	The comparison of performance on <i>ORL</i> database: the first row indicates the number of training sample for each person and the first column indicates the different methods used.	92
4.4	The comparison of performance on <i>UMIST</i> database: the first row indicates the training images used for each person and the first column indicates the different methods used.	96

5.1 The algorithm of Unified Linear Discriminant Analysis 105

5.2 The algorithm of Modified Linear Discriminant Analysis 106

6.1 The algorithm of Random Sub-feature LDA 126

6.2 The algorithm of AdaBoost-LDA 127

6.3 The algorithm of Ensemble Learning for Diversified Fisher Discriminant 130

6.4 Comparison of recognition performance on ORL database: the first
row indicates the number of training sample for each person and the
first column indicates the four proposed methods. 134

6.5 Comparison of recognition performance on YaleB database: the first
row indicates the number of training sample for each person and the
first column indicates the four proposed methods. 134

6.6 Comparison of computational requirements, i.e., CPU Time (s), for
feature extraction on ORL face database (CPU: Pentium IV 2.4GHz,
RAM: 512Mb): the first row indicates the number of training sample
for each person and the first column indicates the methods 135

List of Symbols

$\mathbf{A}_{N \times M}$	An $N \times M$ matrix
$\mathbf{x} \in R^n$	An n-dimensional image vector
$\mathbf{S}_t$	The total covariance matrix
$\mathbf{S}_w$	The within-class covariance matrix
$\mathbf{S}_b$	The between-class covariance matrix
$\mathbf{S}_t^\Phi$	The total covariance matrix in kernel space
$\mathbf{S}_w^\Phi$	The within-class covariance matrix in kernel space
$\mathbf{S}_b^\Phi$	The between-class covariance matrix in kernel space
$\mathbf{W}$	A projection matrix
$\mathbf{W}^\Phi$	A projection matrix in kernel space
$Null(M)$	The null space of matrix M
$Range(M)$	The range space of matrix M
$Rank(M)$	The rank of matrix M
$\ \cdot\ _2$	The 2-norm of a vector
$\ \cdot\ _F$	The F-norm of a matrix
$span\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$	A linear span of $\mathbf{v}_1, \dots, \mathbf{v}_n$
$(\cdot)$	The inner product operation

Chapter 1

Introduction

As one of the most important and successful applications of pattern analysis and machine perception, face recognition has persistently received significant attention for the past twenty years. This attention has become much stronger recently. One of the convincing evidence of such an increasing trend is the emergence of FGR (IEEE International Conference on Face and Gesture Recognition), AVBPA (IAPR International Conference on Audio- and Video-based Biometric Person Authentication), AMFG (IEEE International Workshop on Analysis and Modeling of Faces and Gestures), ICB (IAPR International Conference on Biometrics), FRVT (Face Recognition Vendor Test) [4] and FRGC (Face Recognition Grand Challenge) [5] etc. At least two reasons account for this trend. The first is the wide range of commercial and law enforcement applications for security, and the second is the availability of feasible technologies after twenty years of research.

The general statement of face recognition problem can be formulated as follows: given still or video images of a scene, recognize or identify one or more persons in the scene using a stored database of faces. The solution consists of several sub-tasks in a sequential manner. They are face detection, face normalization, and face recognition. Even though current machine recognition systems have reached a certain level of maturity, they are still limited by the conditions imposed by many

real applications. For example, recognition of face images acquired in an outdoor environment with changes in illumination and/or pose remains a largely unsolved problem. In addition, the limited number of representative training samples available for each subject, which is very common in real-world applications, puts another constraint to the state-of-the-art algorithms. In other words, current systems are still far away from the capability of the human perception system. Therefore, most

Figure 1.1: Illustration of the process of face recognition (face images adopted from CMU PIE [156] and YaleB [158] face image databases).

existing face recognition systems have limited the scope of the problem by dealing primarily with frontal views, neutral expressions, and fixed lighting conditions. To generalize the existing face recognition systems, we look at recognizing faces in varying environment and focus our research on the study and development of methods which can deal with different face expressions, illumination conditions, and pose variations with very limited number of training samples for each subject.

1.1 Motivation

A face image is a visual pattern in the sense that it is a 2D appearance of a 3D face geometry captured by an imaging system. Since any visual appearance of a

3D object is affected by the configuration of an imaging system, without exception, the appearance of the same face is different under different external conditions. An illustration of the imaging system is presented in Figure 1.2. There exist two distinct

Figure 1.2: An illustration of the imaging system.

characteristics of the imaging system, i.e., photometric and geometric properties.

(1) Photometric characteristic is related to the light sources distributed in the scene. The first row of Figure 1.3 shows the face images of one subject captured under varying illumination conditions and the same frontal pose. Numerous models have been proposed to describe the illuminating phenomenon. In addition to the light distribution such as the lighting direction and intensity, an illumination model is in general also related to the object surface material properties, such as Lambertian [123], Phong [72, 73] and Torrance-Sparrow Models [74, 75].

(2) Geometric characteristic is about the camera properties and the relative positioning of the camera and the object. Camera properties include camera intrinsic parameters and camera imaging models. The imaging models widely studied in the computer vision literature are orthographic, scaled orthographic, and perspective

models. The relative positioning of the camera and the object results in pose variation, a key factor determining how the 2D appearances are produced. The second row of Figure 1.3 shows the face images of one subject captured under the same illumination condition but varying poses.

Figure 1.3: Image samples under different imaging conditions (face images are adopted from YaleB [158], ORL [155] and CMU AMP Expression [160] databases).

More generally, the 2D appearances are produced with both illumination and pose change arbitrarily. Figure 1.4 shows the face appearance of the same subject under different illumination and poses. Besides the photometric and geometric characteristics, the face expression is another factor that results in the variation of face appearance. The third row of Figure 1.3 shows the face images of one subject captured under varying facial expressions.

In addition, the number of training samples available is another constraint that prevents the current face recognition algorithms from acquiring higher recognition rate. Generally, there exist a limited number of training samples in realistic face recognition systems. Therefore, solving the Small Sample Size (SSS) problem [56] is another challenge for the design of face recognition algorithms. Subspace analysis methods have achieved promising results and some of them still suffer from the SSS problem, e.g., Linear Discriminant Analysis [56]. Although some methods [38, 39] have been proposed to overcome the SSS problem in LDA, these methods have to discard some useful subspace to make the covariance matrix in question nonsingular.

Figure 1.4: The face images of the same person under different illumination and pose conditions. (face images are adopted from CMU PIE [156] face databases).

1.2 Objectives

Recently, 2D Principal Component Analysis (2DPCA) [9] was proposed and it has demonstrated its superiority over PCA in face recognition. But there are two drawbacks in 2DPCA. Firstly, the theoretical proof why 2DPCA is better than PCA has not been given. Secondly, 2DPCA needs much more coefficients to represent an image than PCA. Thus, it has lower efficiency in recognition and higher storage requirement than PCA. In addition, like PCA, 2DPCA is only good at image representation rather than image discrimination. When there are large pose and illumination variations in face images, the top eigenvectors in 2DPCA do not model the identity information but these external variations and it can be expected that 2DPCA will be inferior to FLD. Therefore, the first objective of this thesis is to overcome the problems in 2DPCA.

Classical FLD projects the data onto a low-dimensional space such that the ratio of the between-class scatter to the within-class scatter is maximized. The optimal projection can be computed by solving a generalized eigenvalue problem. However, the intrinsic limitation of FLD is that its objective function requires the within-class covariance matrix to be non-singular. For many applications such as face

recognition, the scatter matrices in question can be singular due to the SSS problem since the data lie in a high-dimensional space, and the feature dimension far exceeds the number of training samples. To solve the SSS problem in the LDA schemes, various approaches have been proposed [30, 31, 38, 39, 41, 43, 49, 52, 53, 55]. Among them, the two-stage PCA+LDA method [30] have received a lot of attention. In this method, PCA is implemented before LDA. LDA is then performed in the PCA space. Although the scatter matrices in question can be full rank after PCA transformation, the removed subspace may also contain some useful information, and the removal may result in a loss of discriminative information. Although Direct-LDA (D-LDA) [39], Null-space based LDA (N-LDA) [38, 40] and Discriminant Common Vector (DCV) [49] have been proposed to extract the discriminant information from some specific subspaces, they all lose some information in the discarded subspaces. The second objective of this thesis is to construct robust face recognition methods to overcome the problems in conventional LDA-based methods and to kernelize the proposed linear algorithms to improve face recognition accuracy.

1.3 Original Contributions

This thesis deals with face recognition with subspace analysis. In particular, its contributions can be listed as follows:

(1) *Generalized 2D Principal Component Analysis (G2DPCA)*

G2DPCA overcomes the limitations of the recently proposed 2DPCA in three perspectives. Firstly, the essence of 2DPCA is revealed and the theoretical proof why 2DPCA is better than PCA is given. Secondly, as 2DPCA needs much more coefficients to represent an image than PCA, a Bilateral 2DPCA (B2DPCA) is proposed to overcome this problem. Finally, a Kernel 2DPCA (K2DPCA) scheme is introduced and K2DPCA has been shown to be essentially the clustered KPCA method performed on image rows/columns.

Experimental results on ORL, Yale and UMITST face databases show that K2DPCA outperformed 2DPCA by 3%, PCA by 5% and B2DPCA can be more efficient in recognition and face image reconstruction.

(2) *2D Fisher Discriminant Analysis (2DFDA)*

2DFDA is proposed to deal with the SSS problem in Fisher Linear Discriminant analysis (FLD/FDA). Different from FLD, 2DFDA is based on 2D image matrices rather than 1D column vectors so that the image matrix does not need to be transformed into a long vector before feature extraction is carried out. The essence of 2DFDA is revealed by analysis that 2DFDA performed on the image matrices is factually clustered FLD performed on the rows/columns of all the training images. The advantage arising from our algorithm is that the between- and within-class scatter matrices constructed in 2DFDA are both full rank. Our framework, containing Unilateral 2DFDA, Bilateral 2DFDA and Kernel 2DFDA, is applied to face recognition where only few training images exist for each subject. Experimental results on three public databases (ORL, YaleB and UMIST face databases) show that both U2DFDA and B2DFDA outperform the existing linear subspace methods by up to 10% and are comparable to the Kernel Fisher Discriminant Analysis (KFDA), and the Kernel 2DFDA is better than KFDA in recognition rate by up to 4%.

(3) *Discriminant Low-dimensional Subspace Analysis (DLSA)*

DLSA is proposed to deal with the SSS problem in face recognition. In this framework, it is proven that the null space of the total covariance matrix, $\mathbf{S}_t$, has little contribution for recognition purposes. Therefore, a framework of Discriminant Low-dimensional Subspace Analysis (DLSA) is developed in two different ways by discarding the null space of $\mathbf{S}_t$. Two different algorithms are thus proposed. The first algorithm, Unified Linear Discriminant Analysis (ULDA), extracts discriminant information from three subspaces of the range space of $\mathbf{S}_t$. The second algorithm, Modified Linear Discriminant Analysis (MLDA), can avoid the numerical problem in the conventional LDA

by adopting a modified Fisher criterion in a *Deduction* form instead of in the *Quotient* form. Furthermore, the Kernel version of MLDA (KMFDA) is also developed to extract the nonlinear discriminant vectors. Experimental results on a large combined face database have demonstrated that the proposed two linear schemes in this framework can both outperform the state-of-the-art LDA-based algorithms by up to 4%, and the KMFDA is superior to KFDA in terms of recognition accuracy by up to 3%.

(4) *Ensemble Learning for Diversified Fisher Linear Discriminant* (EL-DFLD)

EL-DFLD is proposed to improve the current LDA based face recognition algorithms. Compared with the previous ensemble-based face recognition methods where the ensemble is simply formed by randomly sampling the PCA subspace or/and re-sampling the gallery set according to the random subspace or/and bagging methods, two major distinctions are given in this paper. Firstly, the classifier ensemble in EL-DFLD is composed of a set of diversified component LDA classifiers, which are selected intentionally by computing the diversity between the candidate component classifiers. Secondly, the candidate component classifiers are constructed by coupling a random sub-feature LDA and an adaboost-LDA methods, and it can also be shown that such a coupling scheme will result in more suitable component classifiers so as to increase the generalization performance of EL-DFLD. Experiment results on ORL and YaleB face databases shows that EL-DFLD can achieve higher recognition rate than FLD by 5% to 12% on different databases and outperform 2DFDA by 3%.

1.4 Thesis Organization

The rest of this thesis is organized as follows:

Chapter 2 A review on the existing face recognition algorithms is presented. The characteristic of each algorithm is also provided.

Chapter 3 The framework of Generalized 2D Principal Component Analysis (G2DPCA) is given. This framework consists of B2DPCA and K2DPCA. B2DPCA can give a much more efficient representation of face images. K2DPCA can produce a better generalization ability than 2DPCA.

Chapter 4 The framework of 2D Fisher Discriminant Analysis (2DFDA) is proposed. This framework consists of U2DFDA, B2DFDA and K2DFDA. Extensive experimental results and comparison of 2DFDA with the existing subspace analysis methods are also given.

Chapter 5 The framework of low-dimensional Fisher discriminant subspace analysis is proposed. Unified Linear Discriminant Analysis (ULDA) and Modified Linear Discriminant Analysis (MLDA) are incorporated into this framework. Furthermore, the Kernel Modified Fisher Discriminant Analysis (KMFDA) is proposed to overcome the SSS problem in kernel Fisherface. Experimental results on a large combined face database have shown the performance of the proposed schemes.

Chapter 6 The framework of Ensemble Learning for Diversified Fisher Linear Discriminant is given. The problems existing in the current LDA-based face recognition algorithms are analyzed. A coupling scheme of random sub-feature LDA and AdaBoost-LDA is proposed to overcome these problems.

Chapter 7 This chapter provides the conclusion of this thesis and the proposals for the future research work in subspace analysis for face recognition.

Chapter 2

Face Recognition – A Literature Review

2.1 Introduction

Nowadays, we more and more strongly need user-friendly systems that can secure our assets and protect our privacy without losing our identity in a sea of numbers. At present, one needs a PIN to get cash from an ATM, a password for a computer, a dozen others to access the internet, and so on. Although very reliable methods of biometric personal identification exist, for example, fingerprint analysis and retinal or iris scans, these methods rely on the cooperation of the participants, whereas a personal identification system based on analysis of frontal or profile images of the face is often effective without the participant's cooperation or knowledge. Some of the advantages/disadvantages of different biometrics are described in [7]. Table 2.1 lists some of the applications of face recognition.

Commercial and law enforcement applications of face recognition techniques range from static, controlled-format photographs to uncontrolled video images, posing a wide range of technical challenges and requiring an equally wide range of techniques from image processing, analysis, understanding, and pattern recognition. There-

Table 2.1: Typical Applications of Face Recognition

Areas	Specific Applications
Entertainment	Video game, virtual reality, training programs, Human-robot-interaction, HCI
Smart cards	Drivers' licenses, immigration, national ID, passports, voter registration
Information security	desktop logon, database security, file encryption, Intranet security, internet access
Law enforcement	video surveillance, Shoplifting, suspect tracking and investigation

fore, we will broadly classify face recognition systems into two groups depending on whether they make use of static images or of video. Within either group, significant differences exist depending on the specific application. According to the two major challenges which emerged in the face recognition evaluations, i.e., illumination- and pose-variations in face images, we will give a review of recent face recognition algorithms that can deal with the two problems.

The remainder of this chapter is organized as follows: Section 2.2 gives a short review on PCA, FLD and their kernel counterparts. Section 2.3 and 2.4 reviews the face recognition algorithms based on still images and video sequences. Section 2.5 reviews the pose- and illumination-invariant face recognition algorithms. We draw the concluding remarks in the last section of this chapter.

2.2 Face Recognition Based on Subspace Analysis

Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and their non-linear counterparts, Kernel Principal Component Analysis (KPCA) and Kernel Fisher Discriminant Analysis (KFDA), are very important subspace analysis methods which have been used widely in face recognition area. The following four parts give a brief review of them.

2.2.1 Principal Component Analysis

Principal Component Analysis [57] chooses a dimensionality-reduction linear projection that maximizes the scatter of all the projected samples. Specifically, let us consider a set of N sample images $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ taking values in an n -dimensional image space, and assume that each image belongs to one of c classes $\{X_1, X_2, \dots, X_c\}$. Let us also consider a linear transformation mapping the original n -dimensional image space into an m -dimensional feature space, where $m < n$. The new feature vectors $\mathbf{y}_k \in \mathbf{R}^m$ are defined by the following linear transformation:

$$\mathbf{y}_k = \mathbf{W}^T \mathbf{x}_k \quad k = 1, 2, \dots, N \quad (2.1)$$

where $\mathbf{W} \in \mathbf{R}^{n \times m}$ is a matrix with orthonormal columns.

If the total scatter matrix $\mathbf{S}_t$ is defined as

$$\mathbf{S}_t = \sum_{k=1}^N (\mathbf{x}_k - \bar{\mathbf{m}})(\mathbf{x}_k - \bar{\mathbf{m}})^T \quad (2.2)$$

where $\bar{\mathbf{m}}$ is the mean of the sample images, then after applying the linear transformation $\mathbf{W}^T$, the scatter of the transformed feature vectors $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$ is $\mathbf{W}^T \mathbf{S}_t \mathbf{W}$. In PCA, the projection $\mathbf{W}_{opt}$ is chosen to maximize the determinant of the total scatter matrix of the projected samples as follows,

$$\begin{aligned} \mathbf{W}_{opt} &= \arg \max_{\mathbf{W}} |\mathbf{W}^T \mathbf{S}_t \mathbf{W}| \\ &= [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m] \end{aligned} \quad (2.3)$$

where $\{\mathbf{w}_i | i = 1, 2, \dots, m\}$ is the set of n -dimensional eigenvectors of $\mathbf{S}_t$ corresponding to the m largest eigenvalues. Since these eigenvectors have the same dimension as the original images, they are referred to as Eigenpictures [10] and Eigenfaces [12, 13]. Afterwards, the nearest-neighborhood classifier is adopted in the low-dimensional feature space for classification.

A drawback of this approach is that the scatter being maximized is the sum of the between-class scatter and the within-class scatter. The between-class scatter is useful for recognition while the within-class scatter is unwanted for classification purposes. As stated by Moses et al. [19], much of the variation from one image to the next image is due to illumination changes. Thus, if PCA is presented with images of faces under varying illumination, the projection matrix $\mathbf{W}_{opt}$ will contain principal components (i.e., Eigenfaces) which retain, in the projected feature space, the variation due to lighting variations. Consequently, the points in the projected space will not be well clustered.

2.2.2 Fisher Linear Discriminant

Fisher's Linear Discriminant (FLD) [26] is an example of a class-specific method, in the sense that it tries to "shape" the scatter in order to make it more reliable for classification. This method selects $\mathbf{W}$ in such a way that the ratio of the between-class scatter to the within-class scatter is maximized [8].

Let the between-class scatter matrix be defined as

$$\mathbf{S}_b = \sum_{k=1}^c N_i (\bar{\mathbf{m}}_k - \bar{\mathbf{m}})(\bar{\mathbf{m}}_k - \bar{\mathbf{m}})^T \quad (2.4)$$

and the within-class scatter matrix be defined as

$$\mathbf{S}_w = \sum_{k=1}^c \sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \bar{\mathbf{m}})(\mathbf{x}_k - \bar{\mathbf{m}})^T \quad (2.5)$$

where $\bar{\mathbf{m}}_i$ is the mean of class X_i and N_i is the number of samples in class X_i . If $\mathbf{S}_w$ is nonsingular, the optimal projection $\mathbf{W}_{opt}$ is chosen as the matrix which maximizes the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the pro-

jected samples,

$$\begin{aligned}\mathbf{W}_{opt} &= \arg \max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|} \\ &= [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m]\end{aligned}\quad (2.6)$$

where $\{\mathbf{w}_i | i = 1, 2, \dots, m\}$ is the set of generalized eigenvectors of $\mathbf{S}_b$ and $\mathbf{S}_w$ corresponding to the m largest generalized eigenvalues $\{\lambda_i | i = 1, 2, \dots, m\}$,

$$\mathbf{S}_b \mathbf{w}_i = \lambda_i \mathbf{S}_w \mathbf{w}_i \quad i = 1, 2, \dots, m \quad (2.7)$$

Note that there are at most $c-1$ nonzero generalized eigenvalues, and so an upper bound on m is $c-1$, where c is the number of classes.

Figure 2.1: A comparison of PCA and FLD for a two class problem where data for each class lies near a linear subspace (adopted from [30]).

To illustrate the benefits of class-specific linear projection, a low-dimensional analogue to the classification problem is constructed, where the samples from each class lie near a linear subspace. Figure 2.1 is a comparison of PCA and FLD for a two-

class problem in which the samples from each class are randomly perturbed in a direction perpendicular to a linear subspace. For this example, $N = 20$, $n = 2$, and $m = 1$. So, the samples from each class lie near a line passing through the origin in the 2D feature space. Both PCA and FLD have been used to project the points from 2D down to 1D. Comparing the two projections in the figure, PCA actually smears the classes together so that they are no longer linearly separable in the projected space.

2.2.3 Kernel Principal Component Analysis

Given a set of N (zero mean, unit covariance) samples $\mathbf{x}_k$, $\mathbf{x}_k = [x_{k1}, x_{k2}, \dots, x_{kn}]^T \in R^n$, PCA aims to find the projection directions that maximize the variance of a subspace which is equivalent to finding the eigenvalues from the covariance matrix, $\mathbf{C}$, for eigenvalues $\lambda > 0$ and eigenvectors $\mathbf{w} \in R^n$.

$$\lambda \mathbf{w} = \mathbf{C} \mathbf{w} \quad (2.8)$$

In KPCA, each vector $\mathbf{x}$ is projected from the input space, R^n , to a high dimensional feature space, R^f , by a nonlinear mapping function, $\Phi : R^n \rightarrow R^f, f > n$. Note that the dimensionality of the feature space can be arbitrarily large. In R^f , the corresponding eigenvalue problem is

$$\lambda_i \mathbf{w}_i^\Phi = \mathbf{C}^\Phi \mathbf{w}_i^\Phi \quad (2.9)$$

where $\mathbf{C}^\Phi$ is a covariance matrix. All solutions $\mathbf{w}_i^\Phi$ with $\lambda \neq 0$ lie in the span of $\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \dots, \Phi(\mathbf{x}_N)$, and there exist coefficients $\tilde{\alpha}_i = [\alpha_i^1, \alpha_i^2, \dots, \alpha_i^N]^T$ such that

$$\mathbf{w}_i^\Phi = \sum_{j=1}^N \alpha_i^j \Phi(\mathbf{x}_j) \quad (2.10)$$

Denoting an $N \times N$ matrix K by

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \quad (2.11)$$

Thus, the KPCA problem becomes

$$N\lambda_i \vec{\alpha}_i = K \vec{\alpha}_i \quad (2.12)$$

where $\vec{\alpha}_i$ denotes a column vector with entries $\alpha_i^1, \alpha_i^2, \dots, \alpha_i^N$. The above derivation assumes that all the projected samples are zero-centered in R^f [58].

Note that PCA is a special case of KPCA with first order polynomial kernel. In other words, KPCA is a generalization of classical PCA since different kernels can be used for different nonlinear projections.

We can now project the vectors in R^f to a subspace spanned by the eigenvectors $\mathbf{w}_i^\Phi$. Let $\mathbf{x}$ be a test sample whose projection is $\Phi(\mathbf{x})$ in R^f , the projection of $\Phi(\mathbf{x})$ onto the eigenvectors $\mathbf{w}_i^\Phi$ is the nonlinear principal component corresponding to Φ

$$\mathbf{w}_i^\Phi \cdot \Phi(\mathbf{x}) = \sum_{j=1}^N \alpha_i^j (\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x})) = \sum_{j=1}^N \alpha_i^j k(\mathbf{x}_j, \mathbf{x}) \quad (2.13)$$

In other words, we can extract the first q ($1 \leq q \leq N$) nonlinear principal components using the kernel function without the expensive operation that explicitly projects samples to a high dimensional space R^f .

2.2.4 Kernel Fisher Discriminant Analysis

Similar to the derivations of KPCA, we assume the projected samples $\Phi(\mathbf{x})$ are zero-centered in R^f . We can formulate the equations that use dot products for FLD only. Denoting the between-class and within-class scatter matrices by $\mathbf{S}_w^\Phi$ and $\mathbf{S}_b^\Phi$ and applying FLD in kernel space, we need to find eigenvalues λ and eigenvectors

$\mathbf{w}^\Phi$ of

$$\lambda \mathbf{S}_w^\Phi \mathbf{w}^\Phi = \mathbf{S}_b^\Phi \mathbf{w}^\Phi \quad (2.14)$$

which can be obtained by

$$\mathbf{W}_{opt}^\Phi = \arg \max_{\mathbf{W}^\Phi} \frac{|(\mathbf{W}^\Phi)^T \mathbf{S}_b^\Phi \mathbf{W}^\Phi|}{|(\mathbf{W}^\Phi)^T \mathbf{S}_w^\Phi \mathbf{W}^\Phi|} = [\mathbf{w}_1^\Phi, \mathbf{w}_2^\Phi, \dots, \mathbf{w}_m^\Phi] \quad (2.15)$$

where $\{\mathbf{w}_i^\Phi | i = 1, 2, \dots, m\}$ is the set of generalized eigenvectors corresponding to the m largest generalized eigenvalues $\{\lambda_i | i = 1, 2, \dots, m\}$.

Consider a c -class problem and let the r -th sample of class t and the s -th sample of class u be $\mathbf{x}_{tr}$ and $\mathbf{x}_{us}$ respectively (where class t has l_t samples and class u has l_u samples), the kernel function is defined as follows:

$$(k_{rs})_{tu} = k(\mathbf{x}_{tr}, \mathbf{x}_{us}) = \Phi(\mathbf{x}_{tr}) \cdot \Phi(\mathbf{x}_{us}) \quad (2.16)$$

Let $\mathbf{K}$ be an $N \times N$ matrix defined by the elements $(\mathbf{K}_{tu})_{t=1, \dots, c}^{u=1, \dots, c}$, where $\mathbf{K}_{tu}$ is a matrix composed of dot products in the feature space R^f defined as follows:

$$\mathbf{K} = (\mathbf{K}_{tu})_{t=1, \dots, c}^{u=1, \dots, c} \quad (2.17)$$

where $\mathbf{K}_{tu} = (k_{rs})_{s=1, \dots, l_u}^{r=1, \dots, l_t}$.

Note that $\mathbf{K}_{tu}$ is a $l_t \times l_u$ matrix and $\mathbf{K}$ is an $N \times N$ symmetric matrix. $\mathbf{Z}$ is defined as follows:

$$\mathbf{Z} = (\mathbf{Z}_t)_{t=1, \dots, c} \quad (2.18)$$

where $\mathbf{Z}_t$ is a $l_t \times l_t$ matrix with all terms equal to $\frac{1}{l_t}$. Thus, $\mathbf{S}_b^\Phi$ and $\mathbf{S}_w^\Phi$ are defined as follows:

$$\mathbf{S}_b^\Phi = \sum_{i=1}^c l_i \mathbf{u}_i^\Phi (\mathbf{u}_i^\Phi)^T, \mathbf{S}_w^\Phi = \sum_{i=1}^c \sum_{j=1}^{l_i} \Phi(\mathbf{x}_{ij}) \Phi(\mathbf{x}_{ij})^T \quad (2.19)$$

where $\mathbf{u}_i^\Phi$ is the mean of class i in R^f , l_i is the number of samples in class i . Since

any solution $\mathbf{w}^\Phi \in R^f$ must lie in the span of all the training samples in R^f , i.e.

$$\mathbf{w}^\Phi \in R^f = \sum_{p=1}^c \sum_{q=1}^{l_p} \alpha_{pq} \Phi(\mathbf{x}_{pq}) \quad (2.20)$$

It follows that we can get the solution by solving

$$\lambda \mathbf{K} \mathbf{K} \mathbf{a} = \mathbf{K} \mathbf{Z} \mathbf{K} \mathbf{a} \quad (2.21)$$

Consequently,

$$\mathbf{W}_{opt}^\Phi = \arg \max_{\mathbf{w}^\Phi} \frac{|\mathbf{a}^T \mathbf{K} \mathbf{Z} \mathbf{K} \mathbf{a}|}{|\mathbf{a}^T \mathbf{K} \mathbf{K} \mathbf{a}|} \quad (2.22)$$

Then, $\Phi(\mathbf{x})$ can be projected to a lower dimensional space spanned by eigenvectors $\mathbf{w}^\Phi$ in a way similar to KPCA.

2.3 Face Recognition from Still Images

Many methods of face recognition have been proposed after twenty years of research. Face recognition is such a challenging yet interesting problem that it has attracted researchers with different backgrounds. To have a clear and high-level categorization, we specifically categorize the current face recognition algorithms as follows:

- (1) *Holistic methods.* These methods use the whole face region as the raw input to a recognition system. For example, one of the most widely used representations of the face region is eigenpictures [10,11] which are based on principal component analysis.
- (2) *Feature-based methods.* In these methods, local features such as the eyes, nose, and mouth are first extracted. Their locations and local statistics (geometric and/or appearance) [89,90] are fed into a classifier.

- (3) *Hybrid methods.* As the human perception system does, a machine-based recognition system should use both local features and the whole face region to recognize a face [14, 15]. These methods could potentially offer the best of the two types of methods.

2.3.1 Holistic Approaches

Starting from the successful low-dimensional reconstruction of faces using KL transformation or PCA projection [10, 11], eigenpictures have been one of the major driving forces behind face representation, detection and recognition. It is well known that there exist significant statistical redundancies in natural images. For a limited class of objects such as face images that are normalized with respect to scale, translation, and rotation, the redundancy is even greater [96]. One of the best global compact representations is KL/PCA which de-correlates the outputs.

An advantage of using such representations is their reduced sensitivity to noise. For example, good performance under blurring, partial occlusion and changes in background has been demonstrated in many eigenpicture-based systems, as illustrated in Figure 2.2. This should not come as a surprise since the PCA reconstructed images are much better than the original distorted images in terms of their global appearance, which are demonstrated in Figure 2.3.

Figure 2.2: Electronically modified images which were correctly identified (adopted from [34]).

Figure 2.3: Reconstructed images using 300 PCA projection coefficients for electronically modified images as shown in Figure 2.2 (adopted from [34]).

The first really successful demonstration of machine recognition of faces was made in [12] using eigenfaces for face recognition. Given the eigenfaces, every face in the database can be represented as a vector of weights. The weights are obtained by projecting the image into eigenface components by a simple inner product operation. When a new test image is given, the new image is also represented by its vector of weights. The classification of the test image is carried out by locating the image in the database whose weights are the nearest to the weights of the test image. By using the observation that the projection of a face image and a non-face image are usually different, a method of detecting the presence of a face in a given image is obtained.

Using a probabilistic measure of similarity, the standard eigenface approach was extended to a Bayesian approach [76]. Practically, the major drawback of a Bayesian method is the need to estimate probability distributions in a high-dimensional space from very limited numbers of training samples per class. To avoid this problem, a much simpler two-class problem was created from the multi-class problem by using a similarity measure based on a Bayesian analysis of image differences. Two mutually exclusive classes were defined. Ω_I representing intra-personal variations between multiple images of the same individual, and Ω_E representing extra-personal variations due to differences in identity. Assuming that both classes are Gaussian-distributed, likelihood functions $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ were estimated for a given intensity difference, $\Delta = \mathbf{I}_1 - \mathbf{I}_2$. Given these likelihood functions and using the

Maximum A Posteriori (MAP) rule, two face images are determined to belong to the same individual if $P(\Delta|\Omega_I) > P(\Delta|\Omega_E)$. A large improvement in performance using this technique was reported on the FERET database [2]. In [76], an efficient technique of probability density estimation was proposed by decomposing the input space into the principal subspace F and its orthogonal subspace $\bar{F}$. Covariances only in the principal subspace are estimated for use in the Mahalanobis distance [56]. In Figure 2.4, the so-called dual eigenfaces separately trained on samples from Ω_I and Ω_E are plotted along with the standard eigenfaces. While the extrapersonal eigenfaces appear more similar to the standard eigenfaces than the intrapersonal ones, the intrapersonal eigenfaces represent subtle variations due mostly to expression and lighting, suggesting that they are more critical for identification.

Figure 2.4: Comparison of dual- and standard-eigenface. (a) intrapersonal, (b) extrapersonal and (c) standard (adopted from [76]).

Face recognition systems using LDA/FLD have also been very successful [27, 28, 30, 31]. LDA is carried out via scatter matrix analysis where the between-class and within-class scatter matrices are discussed. In [29], discriminant analysis of eigenfeatures is applied to an image retrieval system to determine not only the generic class (human face vs. non-face objects) but also the individuals within the face

class. Using the tree-structure learning, the eigenspace and LDA projections are recursively applied to smaller and smaller sets of samples. Such recursive partitioning is carried out for every node until the samples assigned to the node belong to a single class. Experiments on this approach were reported in [28].

A comparative performance analysis was carried out in [30]. Four methods were compared in this paper, i.e., (1) a correlation-based method, (2) a variant of the linear subspace method suggested in [17], (3) an eigenface method [12] and (4) a Fisherface method which uses subspace projection prior to LDA projection to avoid the possible singularity in S_w as in [28]. Experiments were performed on a database of 500 images created by Hallinan [18] and a database of 176 images created at Yale [157]. The results of the experiments showed that the Fisherface method performed significantly better than the other three methods. However, no claim was made about the relative performance of these algorithms on larger databases.

Figure 2.5: Different projection bases. First row shows the first five pure LDA basis images. Second row shows the first five subspace LDA basis images. The average face and first four eigenfaces Φ are shown on the third row (adopted from [33]).

To improve the performance of LDA-based systems, a regularized subspace LDA system that unifies PCA and LDA was proposed in [32, 33]. It was later concluded that the global face subspace dimensionality is on the order of 400 for large databases

of 5,000 images [101]. A weighted distance metric in the projection space z was used to improve the performance [34]. Weighted metrics have also been used in the pure LDA approach [27] and the so-called enhanced FLD (EFM) approach [82]. Finally, LDA was regularized by modifying the $\mathbf{S}_w$ matrix to $\mathbf{S}_w + \delta \mathbf{I}$, where δ is a small positive number. The regularization solves the numerical problem when $\mathbf{S}_w$ is close to being singular. In the extreme case where only one sample per class is available, this regularization transforms the LDA problem into a standard PCA problem with $\mathbf{S}_b$ being the covariance matrix $\mathbf{C}$.

An Evolution Pursuit (EP) based adaptive representation and its application to face recognition were presented in [81]. In analogy to projection pursuit methods, EP seeks to learn an optimal basis for the dual purpose of data compression and pattern classification. In order to increase the generalization ability of EP, a balance is sought between minimizing the empirical risk encountered during training and narrowing the confidence interval for reducing the guaranteed risk during future testing on unseen data [102]. Toward that end, EP implements strategies characteristic of Genetic Algorithms (GAs) for searching the space of possible solutions to determine the optimal basis. EP starts by projecting the original data into a lower-dimensional whitened PCA space. Directed random rotations of the basis vectors in this space are then searched by GAs where evolution is driven by a fitness function defined in terms of performance accuracy and class separation. The feasibility of this method has been demonstrated for face recognition, where the large number of possible bases requires a greedy search algorithm. The particular face recognition task involves 1107 FERET frontal face images of 369 subjects; there were three frontal images for each subject, two for training and the remaining one for testing. Improved face recognition performance was reported as compared to eigenfaces, and better generalization capability was achieved than Fisherfaces.

Based on the argument that for tasks such as face recognition much of the important information is contained in high-order statistics, it has been proposed [103] to use ICA to extract features for face recognition. Independent component analysis is a

generalization of principal component analysis, which de-correlates the high-order moments of the input in addition to the second-order moments. Two architectures have been proposed for face recognition, which is shown in Figure 2.6. The first is used to find a set of statistically independent source images that can be viewed as independent image features for a given set of training images [104], and the second is used to find image filters that produce statistically independent outputs (a factorial code method) [105]. In both architectures, PCA is used first to reduce

Figure 2.6: Two architectures for ICA. Left: architecture for finding statistically independent basis images. Right: architecture for finding a factorial code (adopted from [103]).

the dimensionality of the original image size (60×50). ICA is performed on the first 200 eigenvectors in the first architecture, and is carried out on the first 200 PCA projection coefficients in the second architecture. The performance improvement of both architectures over eigenfaces has been reported. A FERET subset consisting of 425 individuals was used. All the frontal views (one per class) were used for training and the remaining (up to three) frontal views for testing.

2.3.2 Feature-Based Approaches

Many methods in the structural matching category have been proposed, including many early methods based on geometry of local features [83,84] as well as 1D [88] and pseudo-2D HMM methods [87]. One of the most successful of these systems

Figure 2.7: The bunch graph representation of faces used in elastic graph matching (adopted from [90]).

is the Elastic Bunch Graph Matching (EBGM) system [89, 90] which is based on DLA [91, 92]. Wavelets, especially Gabor wavelets, play a building block role for facial representation in these graph matching methods. A typical local feature representation consists of wavelet coefficients for different scales and rotations based on fixed wavelet bases (which are called jets in [89]). These locally estimated wavelet coefficients are robust to illumination change, translation, distortion, rotation and scaling.

DLAs attempt to solve some of the conceptual problems of conventional artificial neural networks, the most prominent of these being the representation of syntactical relationships in neural networks. DLAs use synaptic plasticity and are able to form sets of neurons grouped into structured graphs while maintaining the advantages of neural systems. Both [91] and [92] used Gabor-based wavelets as the features. DLAs basic mechanism, in addition to the connection parameter T_{ij} between two neurons (i, j) , is a dynamic variable J_{ij} [92]. Only the J -variables play the roles of synaptic weights for signal transmission. The T -parameters merely act to constrain the J -variables and they can be changed slowly by long-term synaptic plasticity. The weights J_{ij} are subject to rapid modification and are controlled by the signal correlations between neurons i and j . Negative signal correlations lead to a decrease and positive signal correlations lead to an increase in J_{ij} . In the absence of any correlation, J_{ij} slowly returns to a resting state, a fixed fraction of T_{ij} . Each stored

image is formed by picking a rectangular grid of points as graph nodes. The grid is appropriately positioned over the image and is stored with each grid point's locally determined jet, and serves to represent the pattern classes. Recognition of a new image takes place by transforming the image into the grid of jets, and matching all stored model graphs to the image. Conformation of the DLA is done by establishing and dynamically modifying links between vertices in the model domain.

The DLA architecture was recently extended to Elastic Bunch Graph Matching [90], as shown in Figure 2.7. This is similar to the graph described above, but a set of jets instead of only a single jet are attached to each node. The set of jets are called the bunch graph representation and shown in Figure 2.7b. To handle the pose variation problem, the pose of the face is first determined using prior class information [93], and the "jet" transformations under pose variation are learned [94]. Systems based on the EBGM approach have been applied to face detection and extraction, pose estimation, gender classification, sketch-image-based recognition, and general object recognition. The success of the EBGM system may be due to its resemblance to the human visual system [106].

2.3.3 Hybrid Approaches

Hybrid approaches use both holistic and local features. For example, the modular eigenfaces approach [14,15] uses both global eigenfaces and local eigenfeatures. The concept of eigenfaces can be extended to eigenfeatures, such as eigeneyes, eigenmouth etc, as shown in Figure 2.8. Using a limited set of images, recognition performance as a function of the number of eigenvectors was measured for eigenfaces only and for the combined representation. For lower-order spaces, the eigenfeatures performed better than the eigenfaces. Whereas when the combined set was used, only marginal improvement was obtained. These experiments support the claim that feature-based mechanisms may be useful when gross variations are present in the input images.

Figure 2.8: Comparison of matching: (a) test views, (b) eigenface matches, (c) eigenfeature matches (adopted from [14]).

It has been argued that practical systems should use a hybrid of PCA and LFA, as shown in Appendix B in [96]. It seems to be better to estimate eigenmodes/eigenfaces that have large eigenvalues, and therefore the hybrid of PCA and LFA is more robust against noise. While for estimating higher-order eigenmodes, it is better to use LFA. To support this point, it was argued in [96] that the leading eigenpictures are global, integrating, or smoothing filters that are efficient in suppressing noise, while the higher-order modes are ripply or differentiating filters that are likely to amplify noise.

A flexible appearance model based method for automatic face recognition was presented in [97]. To identify a face, both shape and gray-level information are modeled and used. For an input image, three types of information, including extracted shape parameters, shape-free image parameters, and local profiles, are used to compute a Mahalanobis distance for classification as illustrated in Figure 2.9. Based on training 10 and testing 13 images for each of 30 individuals, the classification rate was 92% for the 10 normal testing images and 48% for the three difficult images.

The more recent method [98] that we review in this category is based on recent

Figure 2.9: The face recognition scheme based on flexible appearance model (adopted from [97]).

advances in component-based detection/ recognition [99] and 3D morphable models [100]. The basic idea of component-based methods [99] is to decompose a face into a set of facial components such as mouth and eyes that are interconnected by a flexible geometrical model. The motivation for using components is that changes in head pose mainly lead to changes in the positions of facial components which could be accounted for by the flexibility of the geometric model. However, a major drawback of the system is that it needs a large number of training images taken from different viewpoints and under different lighting conditions. To overcome this problem, the 3D morphable face model [100] is applied to generate arbitrary synthetic images under varying pose and illumination. Only three face images, i.e., frontal, semi-profile, profile images, of a person are needed to compute the 3D face model.

2.4 Face Recognition from Video Sequences

A typical video-based face recognition system automatically detects face regions, extracts features from the video, and recognizes facial identity if a face is present.

In surveillance, information security and access control applications, face recognition and identification from a video sequence is an important problem. Face recognition based on video is preferable over using still images since motion helps in recognition of familiar faces when the images are negated, inverted or threshold [107,108]. It was also demonstrated that humans can recognize animated faces better than randomly rearranged images from the same set.

Historically, video face recognition originated from still-image-based techniques. The system automatically detects and segments the face from the video, and then applies still-image face recognition techniques. Many methods reviewed in previous section such as eigenfaces [12], probabilistic eigenfaces [76] and the EBGM method [89,90] belong to this category. An improvement over these methods is to apply tracking in that a virtual frontal view can be synthesized via pose and depth estimation from video. Due to the abundance of frames in a video, another way to improve the recognition rate is the use of "voting" based on the recognition results from each frame. The voting can be deterministic, but probabilistic voting is better in general [110,111]. One drawback of such voting schemes is the expense of computing the deterministic/probabilistic results for each frame.

The next phase of video-based face recognition will be the use of multimodal cues. Since humans routinely use multiple cues to recognize identities, it is expected that a multimodal system will do better than systems based on faces only. More importantly, using multimodal cues offers a comprehensive solution to the task of identification that might not be achievable by using face images alone.

More recently, a third phase of video based face recognition has started. These methods [114,115] coherently exploit both spatial information in each frame and temporal information among different frames. A big difference between these methods and the probabilistic voting methods [110] is the use of representations in a joint temporal and spatial space for identification. We first review systems that apply still-image-based recognition to selected frames, and followed by multimodal systems. Finally, we review systems that use spatial and temporal information

simultaneously. In [117], a fully automatic person authentication system was described which included video break, face detection, and authentication modules. Perfect results were reported on all three sequences, as verified against a database of 20 still face images. An access control system based on person authentication was described in [118]. The system combined motion and facial appearance, which are two complementary visual cues. In order to reliably detect significant motion, spatiotemporal zero crossings computed from six consecutive frames were used. These motions were grouped into moving objects using a clustering algorithm, and Kalman filters were employed to track the grouped objects. An appearance-based face detection scheme using RBF networks was used to confirm the presence of a person. The face detection scheme was "bootstrapped" using motion and object detection to provide an approximate head region. Face tracking based on the RBF network was used to provide feedback to the motion clustering process to help deal with occlusions. Good tracking results were demonstrated. In [110], this work was extended to person authentication using PCA or LDA. It was argued that recognition based on selected frames is not adequate since important information is discarded. Instead, a probabilistic voting scheme was proposed where face identification was carried out continuously.

An appearance model based method for video tracking and enhancing identification was proposed in [119]. The appearance model is a combination of the Active Shape Model (ASM) [120] and the shape-free texture model after warping the face into a mean shape. Unlike [97] which used the two models separately, a combined set of parameters are used for both models. The main contribution was the decomposition of the combined model parameters into an identity subspace and an orthogonal residual subspace using linear discriminant analysis. Figure 2.10 shows an illustration of separating identity and residue.

In [116], a system called Person-Spotter was described. This system is able to capture, track, and recognize a person walking toward or passing a stereo CCD camera. It has several modules such as a head tracker, pre-selector, landmark finder

Figure 2.10: Active appearance model: Top row shows the appearances by varying the most significant identity parameters and bottom row shows the appearances by manipulating residual variation without affecting identity (adopted from [119]).

and identifier. A recognition rate of about 90% was achieved but the size of the database is not known. A multi-modal person recognition system was described in [109]. This system consists of a face recognition module, a speaker identification module, and a classifier fusion module. The face recognition module can detect and compensate for pose variations. The speaker identification module can detect and compensate for changes in the auditory background. The most reliable video frames and audio clips are selected for recognition. 3D information about the head obtained through Structure from Motion (SfM) is used to detect the presence of an actual person as opposed to an image of that person.

In [114], a face verification system based on tracking facial features was presented. The basic idea of this approach is to exploit the temporal information available in a video sequence to improve face recognition. First, the feature points defined by Gabor attributes on a regular 2D grid are tracked. The trajectories of these tracked feature points are then exploited to identify the person presented in a short video sequence. The proposed tracking-for-verification scheme is different from the pure tracking scheme in that one template face from a database of known persons is selected for tracking. For each template with a specific personal ID, tracking can be performed and trajectories can be obtained. Based on the characteristics of these trajectories, identification can be carried out. According to [114], the trajectories

of the same person are more coherent than those of different persons. Such characteristics can also be observed in the posterior probabilities over time by assuming different classes. In other words, the posterior probabilities for the true hypothesis tend to be higher than those for false hypotheses. This, in turn, can be used for identification. Testing results on a small databases of 19 individuals have suggested that performance is favorable over a frame-to-frame matching and voting scheme, especially in the case of large lighting changes. The testing result is based on comparison with alternative hypotheses.

While most face recognition algorithms take still images as probe inputs, a video-based face recognition approach that takes video sequences as inputs has recently been developed [121]. Since the detected face might be moving in the video sequence, one has to deal with uncertainty in tracking as well as in recognition. Rather than resolving these two uncertainties separately, [121] performed simultaneous tracking and recognition of human faces from a video sequence.

In [115], a multi-view based face recognition system was proposed to recognize faces from videos with large pose variations. To address the challenging pose issue, the concept of an identity surface that captures joint spatial and temporal information was used. An identity surface is a hyper-surface formed by projecting all the images of one individual onto the discriminating feature space parameterized on head pose, as shown in Figure 2.11. To characterize the head pose, two angles, yaw and tilt, are used as basis coordinates in the feature space. As plotted in Figure 2.11, the other basis coordinates represent discriminating feature patterns of faces. Based on recovered pose information, a trajectory of the input feature pattern can be constructed. The trajectories of features from known subjects arranged in the same temporal order can be synthesized on their respective identity surfaces. To recognize a face across views over time, the trajectory for the input face is matched to the trajectories synthesized for the known subjects. This approach can be thought of as a generalized version of face recognition based on single images taken at different poses. Experimental results using twelve training sequences, each containing one

Figure 2.11: Identity surface (adopted from [115]).

subject, and new testing sequences of these subjects were reported. Recognition rates were 100% and 93.9%, using 10 and 2 kernel discriminant analysis (KDA) vectors, respectively.

Other techniques have also been used to construct the discriminating basis in the identity surface. Kernel discriminant analysis (KDA) [122] was used to compute a nonlinear discriminating basis, and a dynamic face model is used to extract a shape-and-pose-free facial texture pattern. The multi-view dynamic face model [113] consists of a sparse Point Distribution Model (PDM) [120], a shape- and pose-free texture model, and an affine geometrical model. The 3D shape vector of a face is estimated from a set of 2D face images in different views using landmark points. A face image fitted by the shape model is then warped to the mean shape in a frontal view, yielding a shape-and pose- free texture pattern. When part of a face is invisible in an image due to rotation in depth, the facial texture is recovered from the visible side of the face using the bilateral symmetry of faces. To obtain a low-dimensional statistical model, PCA was applied to the 3D shape patterns and shape-and-pose-free texture patterns separately. To further suppress within-class variations, the shape-and-pose-free texture patterns were further projected into a KDA feature space. Finally, the identity surface can be approximated and

constructed from discrete samples at fixed poses using a piece-wise planar model.

2.5 Pose- and Illumination-invariant Face Recognition Algorithms

2.5.1 Pose-invariant Face Recognition Algorithms

Earlier methods focused on constructing invariant features [90] or synthesizing a prototypical view, commonly frontal view, after a full model is extracted from the input image [97]. Such methods work well for small rotation angles, but they fail when the angle is large. Most proposed methods are based on using large numbers of multi-view samples.

To assess the pose problem more systematically, an attempt has been made to classify pose problems [32, 35]. The basic idea of this analysis is to use a varying-albedo reflectance model to synthesize new images in different poses from a real image, thus providing a tool for simulating the pose problem. More specifically, the 2D-to-2D image transformation under 3D pose change has been studied. The drawback of this approach is the restriction of using a generic 3D model, and no deformation of this 3D shape was carried out although the authors suggested doing so.

One of the earliest examples of the multiview-based approaches is the work of [133] which used a template-based correlation matching scheme. In this work, pose estimation and face recognition were coupled in an iterative loop. For each hypothesized pose, the input image was aligned to database images corresponding to that pose. The alignment was first carried out via a 2D affine transformation based on three key feature points (eyes and nose), and optical flow was then used to refine the alignment of each template. After this step, the correlation scores of all pairs of matching templates were used for recognition. More recently, an illumination-cone

Figure 2.12: The process of constructing the illumination cone. First row: The seven training images. Second row: Images corresponding to the columns of $\bar{\mathbf{B}}$. Third row: Reconstruction up to a GBR transformation. Fourth row: Synthesized images from the illumination cone of the face under novel lighting conditions but fixed pose (adopted from [70]).

based [127] image synthesis method [70] has been proposed to handle both pose and illumination problems in face recognition. It handles illumination variation quite well, but it cannot deal with pose variation. To handle variations due to rotation, it needs to completely resolve the Generalized-Bas-Relief (GBR) ambiguity and then reconstruct the Euclidean 3D shape. Without resolving this ambiguity, images from non-frontal viewpoints synthesized from a GBR reconstruction will differ from a valid image by an affine warp of the image coordinates. To address the GBR ambiguity, the face symmetry is exploited to correct tilt. The fact that the chin and the forehead are at about the same height is the constraint to correct slant. It is also required that the range of heights of the surface be about twice

the distance between the eyes in order to correct scale [70]. They propose a pose-

Figure 2.13: Synthesized images under variable pose and lighting generated from the training images (adopted from [70]).

and illumination-invariant face recognition method based on building illumination cones at each pose for each person. Though conceptually this is a good idea, it is too expensive to implement. Experiments on building illumination cones and on 3D shape reconstruction based on seven training images per class were reported, as shown in Figure 2.12. Figure 2.13 demonstrates the effectiveness of image synthesis under variable pose and lighting after the GBR ambiguity is resolved. Almost perfect recognition results on ten individuals were reported using nine poses and 45 viewing conditions.

The popular eigenface approach to face recognition has been extended to a view-based eigenface method in order to achieve pose-invariant recognition [14]. This method explicitly codes the pose information by constructing an individual eigenface for each pose. Recently, a unified framework called the bilinear model was proposed in [129] that can handle either pure pose variation or class variation.

The image synthesis method in [138] is based on the assumption of linear 3D object classes and the extension of linearity to images that are 2D projections of the 3D objects. To implement this method, a correspondence between images of the input

object and a reference object is established using optical flow. Correspondences between the reference image and other example images having the same pose are also computed. Finally, the correspondence field for the input image is linearly decomposed into the correspondence fields for the examples. Compared to the parallel deformation scheme in [135], this method reduces the need to compute correspondences between images of different poses. On the other hand, parallel deformation was able to preserve some peculiarities of texture that are nonlinear and that could be "erased" by linear methods. This method was extended in [137] to include an additive error term for better synthesis. In [100], a morphable 3D face model consisting of shape and texture was directly matched to single/multiple input images. As a consequence, head orientation, illumination conditions, and other parameters could be free variables subject to optimization.

In [136], a view-based statistical method was proposed based on a small number of 2D statistical models, Active Appearance Model (AAM). It was argued that their method can handle even profile views in which many features are invisible. To deal with such a large pose variation, they needed sample views at 90° (full profile), 45° (quasi-profile), and 0° (frontal view). A key element that is unique to this method is that for each pose, a different set of features is used. Given a single image of a new person, all the models are used to match the image, and estimation of the pose is achieved by choosing the best fit. To synthesize a new view from the input image, the relationship between models at different views are learned. Figure 2.14 demonstrates the synthesis of a virtual view of a novel face using this method. Results of tracking a face across large pose variations and predicting novel views were reported on a limited data-set of about 15 short sequences. Earlier work on multiview-based methods [133] was extended to explore the prior class information that is specific to a face class and can be learned from a set of prototypes. The key idea of these methods is the vectorized representation of the images at each pose; this is similar to view-based AAM. In both methods [133, 135], an optical flow algorithm is used to compute a dense correspondence between the images. To

Figure 2.14: The best fit to a profile model is projected to the frontal model to predict new views (adopted from [136]).

synthesize a virtual view at pose θ_2 of a novel image at pose θ_1 , the flow between these poses of the prototype images is computed and warped to the novel image after the correspondence between the new image and the prototype image at pose θ_1 is computed. A virtual view can be generated by warping the novel image using the warped flow. Figure 2.15 illustrates a particular procedure adopted in [135]. The parallel deformation needed to compute the flow between the prototype image and the novel image. An obvious drawback of this approach is the difficulty of computing flow when the prototype image and novel image are dramatically different. To handle this issue, [135] proposed first sub-sampling the estimated dense flow to locate local features based on prior knowledge about both images, and then matching the local features.

In [36], a unified approach was proposed to solve both the pose and illumination problems. This method is a natural extension of the method proposed in [35] to handle the illumination problem. Using a generic 3D model, they approximately solved the correspondence problem involved in a 3D rotation, and performed an input-to-prototype image computation. Improved recognition results based on subspace LDA [34] were reported on a small database consisting of frontal and quasi-profile images of 115 novel objects.

Figure 2.15: View synthesis by parallel deformation. (A): the prototype flow is measured between the prototype image and the novel image at the same pose. (B): the flow is mapped onto the novel face. (C): the novel face is 2D-warped to the virtual view (adopted from [135]).

2.5.2 Illumination-invariant Face Recognition Algorithms

The illumination problem is illustrated in Figure 2.16 where the same face appears different due to a change in lighting. The changes induced by illumination are often larger than the differences between individuals, causing systems based on comparing images to misclassify input images. This was experimentally observed in [124] using a data-set of 25 individuals.

Many existing systems use heuristic methods to compensate for lighting changes. For example, simple contrast normalization was used to preprocess the detected faces in [76], while normalization in intensity was carried out by first subtracting a best-fit brightness plane and then applying histogram equalization in [125]. In the face eigen-subspace domain, it was suggested and later experimentally verified in [30] that by discarding a few most significant principal components, variations due to lighting can be reduced. However, in order to maintain system performance

Figure 2.16: In each row, the appearance of the same face is different under different lighting conditions (adopted from [157]).

for normally illuminated images, while improving performance for images acquired under changes in illumination, it must be assumed that the first three principal components capture only variations due to lighting. Other heuristic methods based on frontal-face symmetry have also been proposed [32].

In [124], approaches based on image comparison using different image representations and distance measures were evaluated. The image representations used were edge maps, derivatives of the gray level, images filtered with 2D Gabor-like functions, and a representation that combines a log function of the intensity with these representations. The distance measures used were point-wise distance, regional distance, affine-GL (gray level) distance, local affine-GL distance, and log point-wise distance. More details about these methods and about the evaluation database can be found in [124]. It was concluded that none of these representations alone can overcome the image variations due to illumination. A recently proposed image comparison method [126] used a new measure robust to illumination change. The rationale for developing such a method of directly comparing images is the potential difficulty of building a complete representation of an object's possible images as suggested in [127].

Under the assumptions of Lambertian surfaces and no shadowing, a 3D linear illumination subspace for a person was constructed in [16–18, 127, 128] for a fixed

viewpoint, using three aligned faces/images acquired under different lighting conditions. Under ideal assumptions, recognition based on this subspace is illumination-invariant. More recently, an illumination cone has been proposed as an effective method of handling illumination variations, including shadowing and multiple light sources [70, 127]. This method is an extension of the 3D linear subspace method [17, 18] and has the same drawback, requiring at least three aligned training images acquired under different lighting conditions per person. More recently,

Figure 2.17: Testing the invariance of the Quotient-image (Q-image) to varying illumination. Top row: Original images of a novel face taken under five different illuminations. Bottom row: The Q-images corresponding to the novel images, computed with respect to the bootstrap set of ten objects (adopted from [128]).

a method based on quotient images was introduced [128]. Like other class-based methods, this method assumes that the faces of different individuals have the same shape and different textures. The key result in this approach is based on a definition of an illumination invariant signature image which enables an analytic generation of the image space with varying illumination. It was shown that a small database of objects is sufficient for generating the image space with varying illumination of any new object of the class from a single input image of that object.

The factor contributing to the success of using only a small bootstrap set is that the albedo functions occupy only a small subspace. Figure 2.17 demonstrates the invariance of the quotient image against change in illumination conditions. The image synthesis results are shown in Figure 2.18. In model-based approaches, a

Figure 2.18: Image synthesis example. (a) shows the original image and (b) shows the quotient image of (a) from the $N = 10$ bootstrap set. (c) to (e) are the bootstrap set. (f) to (k) are the synthetic images (adopted from [128]).

3D face model is used to synthesize the virtual image from a given image under desired illumination conditions. When the 3D model is unknown, recovering the shape from the images accurately is difficult without using any priors. Shape-from-Shading (SFS) can be used if only one image is available. Stereo or structure from motion can be used when multiple images of the same object are available.

Fortunately, the differences in the 3D shapes of different face objects are not dramatic. This is especially true after the images are aligned and normalized. Recall that this assumption was used in the class-based methods reviewed above. Using a statistical representation of the 3D heads, PCA was suggested as a tool for solving the parametric SFS problem [130]. An eigen-head approximation of a 3D head was obtained after training on about 300 laser-scanned range images of real human heads. The ill-posed SFS problem is thereby transformed into a parametric problem. It has also been demonstrated that such a representation helps to determine the light source. For a new face image, its 3D head can be approximated as a linear combination of eigen-heads and then used to determine the light source.

To address the issue of varying albedo, a direct 2D-to-2D approach was proposed based on the assumption that front-view faces are symmetric and making use of a generic 3D model [35]. Figure 2.19 shows some comparisons between rendered

images obtained using this method and using a local SFS algorithm [131]. Using the Yale and Weizmann databases, significant performance improvements were reported when the prototype images were used in a subspace LDA system in place of the original input images [35].

Figure 2.19: The comparison of image rendering results. The original images are shown in the first column. The second column shows prototype images rendered using [131]. Prototype images rendered using symmetric SFS are shown in the third column. The fourth column shows real images that are close to the prototype images (adopted from [35]).

Recently, a general method of approximating Lambertian reflectance using second-order spherical harmonics has been reported [132]. Assuming Lambertian objects under distant and isotropic lighting, it can be shown that the set of all reflectance functions can be approximated using the surface spherical harmonic expansion. Specifically, they have proved that the accuracy for any light function exceeds 97.97% using a second-order (nine harmonics) approximation. They then extended this analysis to image formation, which is a much more difficult problem due to possible occlusion, shape, and albedo variations. Using their method, an image can be decomposed into so-called harmonic images, which are produced when the object is illuminated by harmonic functions. The nine harmonic images of a face are plotted in Figure 2.20. An interesting comparison was made between the proposed method and the 3D linear illumination subspace methods [17, 18]; the 3D linear methods are just first-order harmonic approximations without the DC components. Assuming

pre-computed object pose and known color albedo/texture, an 86% recognition rate is reported when applying this technique to the task of face recognition using a probe set of 10 people and a gallery set of 42 people.

Figure 2.20: The first nine harmonic images of a face object (from left to right, top to bottom) (adopted from [132]).

2.6 A Summary of Face Databases Used in this Thesis

In this thesis, altogether seven face databases are used to test the proposed methods. They are ORL [155], Yale [157], YaleB [158], CMU PIE [156], UMIST [159], CMU AMP Expression [160], and XM2VTS databases [6]. They are listed in Table 2.2.

2.7 Concluding Remarks

In this chapter, we have presented an extensive survey of machine recognition of human faces. Firstly, we give a brief introduction of the popular linear/nonlinear subspace analysis methods such as PCA, FLD and their kernel versions (KPCA and KFDA). According to the types of media used, we have considered two types of face

Table 2.2: A Summary of Face Databases Used in This Thesis

Database name	Number of subjects	Number of images	Descriptions
ORL	40	400	with pose, expression, scale variations.
Yale	15	165	frontal pose, expression and illumination.
YaleB	10	5760	9 poses x 64 illumination conditions for each one.
CMU PIE	68	41368	13 poses x 43 illumination conditions for each one.
UMIST	20	564	with pose varying from left profile to right profile.
CMU AMP	13	975	frontal pose, large expression variations.
XM2VTS	295	2360	taken over a period of four months, with a 3D model for each one.

recognition tasks, i.e., one from still images and the other from video sequence. We have categorized the methods used for each type, and discussed their characteristics and their pros and cons. In addition to a detailed review of representative work, we have provided summaries of current developments and of challenging issues. We have also identified two important issues in practical face recognition systems, the illumination and the pose problems. We have categorized proposed methods of solving these problems and discussed the pros and cons of these methods. Finally, a summary of face databases used in this thesis is given.

Chapter 3

Generalized 2D Principal Component Analysis

3.1 Introduction

Principal Component Analysis (PCA) is a classical dimension-reduction method for feature extraction and pattern representation, which applies the Karhunen-Löwe transformation to original data set. It has been widely used in many areas such as face recognition, signal processing and data mining. It was Sirovich and Kirby who originally used PCA to represent the human face images [10, 11]. In face recognition, Turk and Pentland proposed the well-known Eigenface [12]. Since then, PCA-based face/object recognition schemes have been investigated broadly. To deal with pose variation problem, Pentland et al. [14] proposed the view-based and modular eigenspaces. Murase and Naya introduced the appearance manifolds [16]. To overcome the illumination variation problem, Shashua [17], Hallinan [18], Epstein [20], Zhao [21], Ramamoorthi [22], Bischof [23] analyzed the ways of modelling the arbitrary illumination condition for PCA-based recognition methods.

Recently, an image projection technique, 2D Principal Component Analysis (2DPCA) [9], is developed for face recognition. Being different from PCA, 2DPCA is based on

the matrix-based data representation model rather than the vector-based data representation model. That is, the image matrix does not need to be transformed into a vector beforehand. Instead, the covariance matrix is constructed directly using the 2D image matrices. Although it achieves better performance than PCA, there remain several shortcomings in 2DPCA. Firstly, it did not clarify explicitly why 2DPCA is better than PCA. Secondly, a unilateral projection scheme is adopted in the original 2DPCA. However, the disadvantage arising in this way is that more coefficients are needed to represent an image in 2DPCA than PCA. Thirdly, 2DPCA is a linear method, which neglects the higher-order statistics among the row/column vectors of the images. However, it is well known that the object/face appearances lie on a nonlinear low-dimensional manifold when there exist pose or/and illumination variations [16]. 2DPCA cannot effectively model such a nonlinearity, and this prevents it from higher recognition rate.

Another dimension-reduction method for face recognition, Fisher Linear Discriminant (FLD) [56], achieves maximum discrimination by projecting the data onto a lower-dimensional vector space such that the ratio of the between-class scatter to the within-class scatter is maximized. In FLD, the optimal projection can be readily computed by solving a generalized eigenvalue problem. However, because of the Small Sample Size problem, the within-class covariance matrix, S_w , is singular so that the numerical problem is introduced in solving the optimal discriminating directions. To solve the singularity problem, several two-stage LDA were proposed [28], [30], [31] and [49].

It should be pointed out that FLD is good at discrimination rather than representation. FLD can generally achieve better performance than PCA when there exist noticeable illumination and pose variations. However, FLD will be inferior to PCA if the illumination and pose variations are not so significant and there are very limited number of training samples for each subject. The reasons are listed as follows: Firstly, when there are large pose and illumination variations in face images, the top eigenvectors in PCA do not model identity information but these external vari-

ations. Secondly, the null space of $\mathbf{S}_w$ is discarded in FLD. The rank of the null space of $\mathbf{S}_w$ is $(C - 1)$ in general. When the number of training samples for each subject is small, the rank of null space of $\mathbf{S}_w$ is comparable to the rank of range space of $\mathbf{S}_w$. Therefore, discarding the whole null space of $\mathbf{S}_w$ is equivalent to losing a large quantity of discriminant information. However, with the number of training samples increasing, the rank of null space of $\mathbf{S}_w$ is much smaller than the rank of range space of $\mathbf{S}_w$ and discarding the whole null space of $\mathbf{S}_w$ will lose relatively little useful information.

In this chapter, a framework of Generalized 2D Principal Component Analysis (G2DPCA) is proposed to extend the original 2DPCA in the following three ways. Firstly, the essence of 2DPCA is revealed to explain why 2DPCA can achieve better performance than PCA in face recognition. Secondly, a Bilateral 2DPCA, called B2DPCA, is introduced to simultaneously construct two subspaces to encode the row and column vectors of the image matrices respectively. In this way, B2DPCA can achieve higher compression rate and efficiency than 2DPCA and PCA. Thirdly, a Kernel 2DPCA (K2DPCA) is investigated to remedy the shortcoming of 2DPCA in modelling the nonlinear manifold of face images.

The remainder of this chapter is organized as follows: 2DPCA algorithm is reviewed in Section 3.2. The essence of 2DPCA and the relationship between 2DPCA and PCA are revealed in Section 3.3. B2DPCA algorithm and its application in image reconstruction/recognition are given in Section 3.4. The Kernel 2DPCA is introduced in Section 3.5. Experimental results are presented in Section 3.6. We draw the concluding remarks in the last section of this chapter.

3.2 2D Principal Component Analysis

Let $\mathbf{x}$ be an n -dimensional unitary column vector. The idea is to project an image $\mathbf{A}$, an $m \times n$ matrix, onto $\mathbf{x}$ by $\mathbf{y} = \mathbf{Ax}$. To determine the optimal projection

vector $\mathbf{x}$, the total scatter of the projected samples, $\mathbf{S}_x$, is used to measure the goodness of $\mathbf{x}$. $\mathbf{S}_x = \mathbf{x}^T E\{[\mathbf{A} - E(\mathbf{A})]^T [\mathbf{A} - E(\mathbf{A})]\} \mathbf{x} = \mathbf{x}^T \mathbf{S}_A \mathbf{x}$, where $\mathbf{S}_A = E\{[\mathbf{A} - E(\mathbf{A})]^T [\mathbf{A} - E(\mathbf{A})]\}$, called the image covariance matrix. Suppose that there are totally M training samples $\{\mathbf{A}_i\}, i = 1, 2, \dots, M$, and the average image is denoted by $\bar{\mathbf{A}}$, then $\mathbf{S}_A = \frac{1}{M} \sum_{i=1}^M [\mathbf{A}_i - \bar{\mathbf{A}}]^T [\mathbf{A}_i - \bar{\mathbf{A}}]$. The optimal projection direction, $\mathbf{x}_{opt}$, is the eigenvector of $\mathbf{S}_A$ corresponding to the largest eigenvalue. Usually a set of orthonormal projection directions, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d$, are selected and these projection directions are the orthonormal eigenvectors of $\mathbf{S}_A$ corresponding to the first d largest eigenvalues. For a given image $\mathbf{A}$, let $\mathbf{y}_k = \mathbf{A} \mathbf{x}_k, k = 1, 2, \dots, d$. A set of projected feature vectors $\mathbf{y}_k$, the *principal components (vectors)* of $\mathbf{A}$, are obtained. Then the *feature matrix* of $\mathbf{A}$ is formed as $\mathbf{B} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_d]$. The nearest-neighborhood classifier is adopted for classification. The distance between two arbitrary *feature matrices*, $\mathbf{B}_i$ and $\mathbf{B}_j$, is defined as $d(\mathbf{B}_i, \mathbf{B}_j) = \sum_{k=1}^d \|y_k^i - y_k^j\|_2$, where $\|y_k^i - y_k^j\|_2$ is the Euclidean distance between y_k^i and y_k^j .

3.3 The Essence of 2D Principal Component Analysis

The essence of the 2DPCA is not discussed in [9]. However, it is important to investigate the essence for understanding the advantages of the 2DPCA over PCA. Apparently, the newly defined covariance matrix in 2DPCA should be more physically meaningful in the matrix space than in the vector space. However, Theorem 3.1 will give another perspective to make the 2DPCA physically meaningful even in vector space.

Theorem 3.1: 2DPCA performed on the 2D images is essentially a clustered PCA performed on all the rows of all the images if each row is viewed as a computational unit.

Proof: Let $\mathbf{A}_i$ be the i -th training sample, $\mathbf{A}_i^j$ be the j -th row of $\mathbf{A}_i$. Let $E(\mathbf{A})$ be

the mean of all the training samples, $E(\mathbf{A})^j$ be the j -th row of $E(\mathbf{A})$. Let $\hat{\mathbf{A}}_i$ be the zero-centered $\mathbf{A}_i$ and $\hat{\mathbf{A}}_i^j$ be the zero-centered $\mathbf{A}_i^j$, where $\hat{\mathbf{A}}_i = \mathbf{A}_i - E(\mathbf{A})$ and $\hat{\mathbf{A}}_i^j = \mathbf{A}_i^j - E(\mathbf{A})^j$.

Because of the limited number of available samples in specific applications, $\mathbf{S}_A$ is often estimated by:

$$\mathbf{S}_A = \frac{1}{M} \sum_{i=1}^M [\mathbf{A}_i - E(\mathbf{A})]^T [\mathbf{A}_i - E(\mathbf{A})] \quad (3.1)$$

It can also be written as,

$$\mathbf{S}_A = \frac{1}{M} \Psi \Psi^T \quad (3.2)$$

where

$$\Psi = [[\mathbf{A}_1 - E(\mathbf{A})]^T, \dots, [\mathbf{A}_M - E(\mathbf{A})]^T] \quad (3.3)$$

or

$$\Psi = [[(\hat{\mathbf{A}}_1^1)^T, \dots, (\hat{\mathbf{A}}_1^m)^T], \dots, [(\hat{\mathbf{A}}_M^1)^T, \dots, (\hat{\mathbf{A}}_M^m)^T]] \quad (3.4)$$

Therefore, if each row of the image matrix is viewed as a new computational unit, $\mathbf{S}_A$ can be viewed as the covariance matrix evaluated using a set of sub-covariance matrices which are constructed using corresponding rows of all the training samples. Thus, 2DPCA is a clustered PCA (each row of the image matrix is a unit). In 2DPCA, the maximization of $\mathbf{S}_x$ is equal to maximize $\mathbf{x}^T \Psi \Psi^T \mathbf{x}$. This translates into the eigen-analysis of $\Psi \Psi^T$:

$$\lambda_i \mathbf{x}_i = \Psi \Psi^T \mathbf{x}_i \quad (3.5)$$

Hence, 2DPCA performed on the image matrices is essentially the clustered PCA performed on the rows of all the images. $\square$

As a result, we have concluded the following advantages of 2DPCA over PCA. Firstly, as the dimension of the row of each image is much smaller than that of the

entire image vector, the dilemma of *Curse of Dimensionality* diminishes. Secondly, the distance function adopted in the classification criterion in 2DPCA is a global combination of all the local Eigen-feature distances.

3.4 Bilateral 2D Principal Component Analysis

As mentioned in Section 3.1, 2DPCA is a unilateral-projection based scheme. However, as we have described, 2DPCA is actually clustered PCA performed on the rows of all the images. Therefore, the unilateral scheme will lose those information embedded in the columns of the images. In addition, a disadvantage resulting from the unilateral-projection scheme is that more coefficients are needed to represent an image. Consider this, a Bilateral 2DPCA (B2DPCA) is proposed in this section. Compared with the standard 2DPCA, it can remove the redundancies both among the rows and columns of the images, and be consequently able to reduce the number of coefficients for representing an image. Additionally, it encodes completely the spatial information that is beneficial for classification.

3.4.1 The Algorithm

Let $\mathbf{U} \in \mathcal{R}^m \times \mathcal{R}^l$ and $\mathbf{V} \in \mathcal{R}^n \times \mathcal{R}^r$ be the left- and right-multiplying projection matrix respectively. It is assumed that all the samples are all zero-centered in the later sections. For an $m \times n$ image $\mathbf{A}_i$ and an $l \times r$ projected image $\mathbf{B}_i$, the bilateral projection is formulated as follows:

$$\mathbf{B}_i = \mathbf{U}^T \mathbf{A}_i \mathbf{V} \quad (3.6)$$

where $\mathbf{B}_i$ is the extracted feature matrix for image $\mathbf{A}_i$.

The common optimal projection matrices, $\mathbf{U}_{Opt}$ and $\mathbf{V}_{Opt}$ in Equation (3.6) can be computed by solving the following minimization problem such that $\mathbf{U}_{Opt} \mathbf{B}_i \mathbf{V}_{Opt}^T$

gives the best approximation of $\mathbf{A}_i, i = 1, \dots, M$:

$$[\mathbf{U}_{Opt}, \mathbf{V}_{Opt}] = \arg \min \sum_{i=1}^M \|\mathbf{A}_i - \mathbf{U}\mathbf{B}_i\mathbf{V}^T\|_F^2 \quad (3.7)$$

where M is the number of data samples and $\|\cdot\|_F$ is the Frobenius norm of a matrix.

Or $\mathbf{U}_{Opt}$ and $\mathbf{V}_{Opt}$ can be obtained by maximizing $\sum_{i=1}^M \|\mathbf{U}^T \mathbf{A}_i \mathbf{V}\|_F^2$.

Lemma 3.1: Minimizing Equation (3.7) is equivalent to the maximization of $\sum_{i=1}^M \|\mathbf{U}^T \mathbf{A}_i \mathbf{V}\|_F^2$.

Proof: Let $\nabla = \sum_{i=1}^M \|\mathbf{A}_i - \mathbf{U}\mathbf{B}_i\mathbf{V}^T\|_F^2$. According to the property of *trace* of matrix, we have

$$\begin{aligned} \nabla &= \sum_{i=1}^M \text{tr}((\mathbf{A}_i - \mathbf{U}\mathbf{B}_i\mathbf{V}^T)(\mathbf{A}_i - \mathbf{U}\mathbf{B}_i\mathbf{V}^T)^T) \\ &= \sum_{i=1}^M \text{tr}(\mathbf{A}_i\mathbf{A}_i^T) + \text{tr}(\mathbf{U}\mathbf{B}_i\mathbf{V}^T\mathbf{V}\mathbf{B}_i^T\mathbf{U}^T) - 2\text{tr}(\mathbf{U}\mathbf{B}_i\mathbf{V}^T\mathbf{A}_i^T) \\ &= \sum_{i=1}^M \text{tr}(\mathbf{A}_i\mathbf{A}_i^T) + \text{tr}(\mathbf{U}\mathbf{B}_i\mathbf{B}_i^T\mathbf{U}^T) - 2\sum_{i=1}^M \text{tr}(\mathbf{U}\mathbf{B}_i\mathbf{V}^T\mathbf{A}_i^T) \\ &= \sum_{i=1}^M \text{tr}(\mathbf{A}_i\mathbf{A}_i^T) + \sum_{i=1}^M \text{tr}(\mathbf{B}_i^T\mathbf{U}^T\mathbf{U}\mathbf{B}_i) - 2\text{tr}(\mathbf{U}\mathbf{B}_i\mathbf{V}^T\mathbf{A}_i^T) \\ &= \sum_{i=1}^M \{\text{tr}(\mathbf{A}_i\mathbf{A}_i^T) + \text{tr}(\mathbf{B}_i^T\mathbf{B}_i) - 2\text{tr}(\mathbf{U}\mathbf{B}_i\mathbf{V}^T\mathbf{A}_i^T)\} \\ &= \sum_{i=1}^M \{\text{tr}(\mathbf{A}_i\mathbf{A}_i^T) + \text{tr}(\mathbf{B}_i\mathbf{B}_i^T) - 2\text{tr}(\mathbf{U}\mathbf{B}_i\mathbf{V}^T\mathbf{A}_i^T)\} \end{aligned} \quad (3.8)$$

where the second term derives from the facts that: Firstly, both $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns. Secondly $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ for any two matrices.

Since the first term is a constant, the minimization of Equation (3.7) is equivalent to minimizing:

$$\mathbf{J} = \sum_{i=1}^M \{\text{tr}(\mathbf{B}_i\mathbf{B}_i^T) - 2\text{tr}(\mathbf{U}\mathbf{B}_i\mathbf{V}^T\mathbf{A}_i^T)\} \quad (3.9)$$

Let,

$$\frac{\partial \mathbf{J}}{\partial \mathbf{B}_i} = 2 \sum_{i=1}^M \{\mathbf{B}_i - \mathbf{U}^T \mathbf{A}_i \mathbf{V}\} = 0 \quad (3.10)$$

Therefore, only if $\mathbf{B}_i = \mathbf{U}^T \mathbf{A}_i \mathbf{V}$, the minimum value of $\mathbf{J}$ can be achieved. We substitute $\mathbf{B}_i$ in Equation (3.8) by $\mathbf{U}^T \mathbf{A}_i \mathbf{V}$:

$$\begin{aligned} \nabla &= \sum_{i=1}^M \{tr(\mathbf{A}_i \mathbf{A}_i^T) - tr(\mathbf{B}_i \mathbf{B}_i^T)\} \\ &= \sum_{i=1}^M \|\mathbf{A}_i\|_F^2 - \sum_{i=1}^M \|\mathbf{B}_i\|_F^2 \\ &= \sum_{i=1}^M \|\mathbf{A}_i\|_F^2 - \sum_{i=1}^M \|\mathbf{U}^T \mathbf{A}_i \mathbf{V}\|_F^2 \end{aligned} \quad (3.11)$$

where the first term is a constant. Therefore, minimization of Equation (3.7) is equivalent to the maximization of Equation (3.12). The solutions that maximize Equation (3.12) are the optimal ones.

$$\delta = \sum_{i=1}^M \|\mathbf{U}^T \mathbf{A}_i \mathbf{V}\|_F^2 \quad (3.12)$$

□

Given the data set $\mathbf{A}_i \in \mathcal{R}^m \times \mathcal{R}^n, i = 1, \dots, M$, the covariance matrix of the projected samples is defined as:

$$\mathbf{C} = \frac{1}{M} \sum_{i=1}^M \mathbf{B}_i^T \mathbf{B}_i \quad (3.13)$$

where $\mathbf{B}_i$ is defined in Equation 3.6. By replacing $\mathbf{B}_i$ with $\mathbf{U}^T \mathbf{A}_i \mathbf{V}$, it translates into:

$$\mathbf{C} = \frac{1}{M} \sum_{i=1}^M (\mathbf{U}^T \mathbf{A}_i \mathbf{V})^T (\mathbf{U}^T \mathbf{A}_i \mathbf{V}) \quad (3.14)$$

and it is trivial to check that $tr(\mathbf{C}) = \frac{1}{M} \sum_{i=1}^M \|\mathbf{U}^T \mathbf{A}_i \mathbf{V}\|_F^2$.

In this regard, maximizing the trace of the covariance matrix of the projected samples is equivalent to maximizing $\sum_{i=1}^M \|\mathbf{U}^T \mathbf{A}_i \mathbf{V}\|_F^2$, while maximizing $\sum_{i=1}^M \|\mathbf{U}^T \mathbf{A}_i \mathbf{V}\|_F^2$ has been shown to be equivalent to minimizing $\sum_{i=1}^M \|\mathbf{A}_i - \mathbf{U} \mathbf{B}_i \mathbf{V}^T\|_F^2$ and optimally reconstructing (approximating) the images. Therefore, the proposed bilateral-projection scheme is consistent with the principle of PCA and 2DPCA, and it can be viewed as a generalized 2DPCA, i.e., the standard 2DPCA is a special form of the bilateral 2DPCA.

To our knowledge, there is no close-form solution to maximize $\sum_{i=1}^M \|\mathbf{U}^T \mathbf{A}_i \mathbf{V}\|_F^2$ because $\mathbf{C} = \frac{1}{M} \sum_{i=1}^M \mathbf{V}^T \mathbf{A}_i^T \mathbf{U} \mathbf{U}^T \mathbf{A}_i \mathbf{V}$ and there is no direct method for the eigen-decomposition of such a coupled covariance matrix. Considering this, an iterative algorithm is proposed to compute the $\mathbf{U}_{Opt}$ and $\mathbf{V}_{Opt}$. Before we give the details of the iterative algorithm, we have the following two Lemmas.

Theorem 3.2 : Given $\mathbf{U}_{Opt}$, $\mathbf{V}_{Opt}$ can be obtained as the matrix formed by the first r eigenvectors corresponding to the first r largest eigenvalues of $\mathbf{C}_v = \frac{1}{M} \sum_{i=1}^M \mathbf{A}_i^T \mathbf{U}_{Opt} \mathbf{U}_{Opt}^T \mathbf{A}_i$.

Proof: Since $\mathbf{U}_{Opt}$ and $\mathbf{V}_{Opt}$ maximize $tr(\frac{1}{M} \sum_{i=1}^M \mathbf{V}^T \mathbf{A}_i^T \mathbf{U} \mathbf{U}^T \mathbf{A}_i \mathbf{V})$. If $\mathbf{U}_{Opt}$ is known, $tr(\mathbf{C}) = tr(\frac{1}{M} \sum_{i=1}^M \mathbf{V}^T \mathbf{A}_i^T \mathbf{U}_{Opt} \mathbf{U}_{Opt}^T \mathbf{A}_i \mathbf{V}) = tr(\mathbf{V}^T \mathbf{C}_v \mathbf{V})$.

Therefore, the maximization of $tr(\mathbf{C})$ equals to solve the first r eigenvectors of $\mathbf{C}_v$ corresponding to the first r largest eigenvalues. $\square$

Theorem 3.3: Given $\mathbf{V}_{Opt}$, $\mathbf{U}_{Opt}$ can be obtained as the matrix formed by the first l eigenvectors corresponding to the first l largest eigenvalues of $\mathbf{C}_u = \frac{1}{M} \sum_{i=1}^M \mathbf{A}_i \mathbf{V}_{Opt} \mathbf{V}_{Opt}^T$.

The proof of Theorem 3.3 is similar to that of Theorem 3.2. $\square$

By Theorems 3.2 and 3.3, the detailed iterative scheme to compute the $\mathbf{U}_{Opt}$ and $\mathbf{V}_{Opt}$ is listed in Table 3.1.

Theoretically, the obtained solutions are locally optimal because the solutions are dependent on the initialized $\mathbf{U}_0$. By extensive experiments, $\mathbf{U}_0 = \mathbf{I}_m$, a setting we adopted, will produce excellent results. Another issue deserving attention is the

Table 3.1: The algorithm for computing $\mathbf{U}_{Opt}$ and $\mathbf{V}_{Opt}$

S_1 .	Initialize $\mathbf{U}$, $\mathbf{U} = \mathbf{U}_0$, $i = 0$ and $\mathcal{E}(0)$ is a very large positive number
S_2 .	While not convergent
S_3 .	Compute $\mathbf{C}_v$ and the eigenvectors $\{e_j^V\}_{j=1}^r$ corresponding to its r top eigenvalues, then $\mathbf{V}_i \leftarrow [e_1^V, \dots, e_r^V]$
S_4 .	Compute $\mathbf{C}_u$ and the eigenvectors $\{e_j^U\}_{j=1}^l$ corresponding to its l top eigenvalues, then $\mathbf{U}_i \leftarrow [e_1^U, \dots, e_l^U]$
S_5 .	$i \leftarrow i + 1$
S_6 .	$\mathbf{V}_{Opt} \leftarrow \mathbf{V}_{i-1}$ and $\mathbf{U}_{Opt} \leftarrow \mathbf{U}_{i-1}$
S_7 .	Feature extraction: $\mathbf{B}_k = \mathbf{U}_{Opt}^T \mathbf{A}_k \mathbf{V}_{Opt}$, $k=1, \dots, M$
S_8 .	Compute the reconstruction error: $\mathcal{E}(i) = \frac{1}{M} \sum_{k=1}^M \ \mathbf{A}_k - \mathbf{U}_{Opt} \mathbf{B}_k \mathbf{V}_{Opt}^T\ _F$
S_9 .	Compute the criterion for convergence: $\frac{\mathcal{E}(i-1) - \mathcal{E}(i)}{\mathcal{E}(i-1)}$
S_{10} .	End While

---

convergence of this algorithm. We defined it to be the mean reconstruction error, i.e.

$$\mathcal{E} = \frac{1}{M} \sum_{i=1}^M \|\mathbf{A}_i - \mathbf{U} \mathbf{B}_i \mathbf{V}^T\|_F \quad (3.15)$$

We thus use the relative reduction of the $\mathcal{E}$ value to check the convergence of the iterative algorithm. More specifically, let $\mathcal{E}(i)$ and $\mathcal{E}(i-1)$ be the error at the i -th and $(i-1)$ -th iteration respectively. The convergence of this algorithm can be judged by whether it can satisfy the following inequity.

$$\frac{\mathcal{E}(i-1) - \mathcal{E}(i)}{\mathcal{E}(i-1)} \leq \mu \quad (3.16)$$

where μ is a small positive number. Our experiments in the later section will show that the iterative algorithm usually converges in two iterations.

3.4.2 Image Representation Using B2DPCA

Since we have obtained the common optimal projection matrices, $\mathbf{U}_{Opt}^T \in \mathcal{R}^m \times \mathcal{R}^l$ and $\mathbf{V}_{Opt} \in \mathcal{R}^n \times \mathcal{R}^r$, for any image $\mathbf{A}_i \in \mathcal{R}^m \times \mathcal{R}^n$, its feature matrix $\mathbf{B}_i \in \mathcal{R}^l \times \mathcal{R}^r = \mathbf{U}_{Opt}^T \mathbf{A}_i \mathbf{V}_{Opt}$. Therefore, $\mathbf{B}_i$ is the coefficient matrix that can be used to reconstruct

the image $\mathbf{A}_i$ by $\hat{\mathbf{A}}_i = \mathbf{U}_{Opt} \mathbf{B}_i \mathbf{V}_{Opt}^T$.

3.4.3 Face Recognition Using B2DPCA

After a transformation by B2DPCA, a feature matrix, $\mathbf{B}_i \in \mathcal{R}^l \times \mathcal{R}^r$, is obtained for each image. The nearest-neighbor classifier is used for classification. Here, the distance between two arbitrary feature matrices, $\mathbf{B}_i = [\mathbf{b}_i^1, \dots, \mathbf{b}_i^r]$ and $\mathbf{B}_j = [\mathbf{b}_j^1, \dots, \mathbf{b}_j^r]$, is defined by

$$\text{dist}(\mathbf{B}_i, \mathbf{B}_j) = \sum_{k=1}^r \|\mathbf{b}_i^k - \mathbf{b}_j^k\|_2 \quad (3.17)$$

where $\|\mathbf{b}_i^k - \mathbf{b}_j^k\|_2$ denotes the Euclidean distance between the two feature vectors $\mathbf{b}_i^k$ and $\mathbf{b}_j^k$.

Suppose that the training samples are $\mathbf{B}_1, \dots, \mathbf{B}_M$ (where M is the total number of training samples), and that each of these samples is assigned a given identity (class) C_k . Given a test sample $\mathbf{B}$, if $\text{dist}(\mathbf{B}, \mathbf{B}_l) = \min_{i=1}^M \text{dist}(\mathbf{B}, \mathbf{B}_i)$, and $\mathbf{B}_l \in C_t$, then the resulting decision is $\mathbf{B} \in C_t$.

3.5 Kernel 2D Principal Component Analysis

Kernel Principal Component Analysis (KPCA) [58] is the generalized Principal Component Analysis by mapping the input data onto a higher- or infinite-dimensional space and utilizing the kernel trick to simplify the computations. The kernel method is able to capture higher order statistical dependencies among the input data. KPCA has been applied to face recognition [59] and it has demonstrated better recognition performance than PCA. Likewise, FLD is also extended to the kernel space in [60], [65], [66]. Similarly, the kernelization of 2DPCA inspires from the fact that we are more interested in the *principal components* of the features that are nonlinearly related to the input image matrix rather than those directly extracted from

the original input image matrix. Inspired by the advantage of 2DPCA over PCA, it is desirable to keep the integral 2D structure of the image matrix rather than to disintegrate and then transform it to a long vector. Therefore, those features nonlinearly related to the input image matrix can be obtained by exploring the higher-order correlations among row/column vectors. Similar to KPCA, a nonlinear mapping without explicit function is performed. Different from KPCA, this mapping is performed on each row vector of all the training and test image matrices, i.e. let $\Phi : \mathbf{R}^t \rightarrow \mathbf{R}^f, f > t$, be a nonlinearly mapping on each row of the image, where t is the length of the rows of an image and f can be arbitrarily large. The dot product in the feature space of $\mathbf{R}^f$ can be conveniently calculated via a pre-defined kernel function, such as the commonly used Gaussian RBF kernel function.

It is assumed that all the mapped data are zero-centered by the method in [58]. Let $\hat{\Phi}(\mathbf{A}_i)$ be the i -th mapped image in which $\hat{\Phi}(\mathbf{A}_i^j)$ be the j -th zero-centered row vector of it. The covariance matrix $\mathbf{C}^\Phi$ in $\mathbf{R}^f$:

$$\mathbf{C}^\Phi = \frac{1}{M} \sum_{i=1}^M \hat{\Phi}(\mathbf{A}_i)^T \hat{\Phi}(\mathbf{A}_i) \quad (3.18)$$

where $\hat{\Phi}(\mathbf{A}_i) = [\hat{\Phi}(\mathbf{A}_i^1)^T, \hat{\Phi}(\mathbf{A}_i^2)^T, \dots, \hat{\Phi}(\mathbf{A}_i^m)^T]^T$ and m is the number of row vectors. Therefore, if $\mathbf{R}^f$ is infinite-dimensional, $\mathbf{C}^\Phi$ is $\text{inf} \times \text{inf}$ in size. It is intractable to directly calculate the eigenvalues, λ_i , and the eigenvectors, $\mathbf{v}_i$, that satisfy

$$\lambda_i \mathbf{v}_i = \mathbf{C}^\Phi \mathbf{v}_i \quad (3.19)$$

However, K2DPCA can be implemented using KPCA according to the following theorem.

Theorem 3.4: K2DPCA on the images is essentially the clustered KPCA performed on the rows of all the training image matrices if each row is viewed as a computational unit.

Proof: From Equation (3.18) and Equation (3.19), we have $\mathbf{v}_i = \frac{1}{\lambda_i} \mathbf{C}^\Phi \mathbf{v}_i$.

$$\mathbf{v}_i = \frac{1}{\lambda_i} \left[\frac{1}{M} \sum_{k=1}^M \hat{\Phi}(\mathbf{A}_k)^T \hat{\Phi}(\mathbf{A}_k) \right] \mathbf{v}_i \quad (3.20)$$

Another form of $\mathbf{C}^\Phi$ is

$$\mathbf{C}^\Phi = \frac{1}{M} \Psi^\Phi (\Psi^\Phi)^T \quad (3.21)$$

where

$$\Psi^\Phi = [[\hat{\Phi}(\mathbf{A}_1^1)^T, \dots, \hat{\Phi}(\mathbf{A}_1^m)^T], \dots, [\hat{\Phi}(\mathbf{A}_M^1)^T, \dots, \hat{\Phi}(\mathbf{A}_M^m)^T]] \quad (3.22)$$

From Equation (3.20), Equation (3.21) and Equation (3.22), we have,

$$\mathbf{v}_i = \frac{1}{\lambda_i M} \Psi^\Phi \mathbf{a}_i \quad (3.23)$$

where $\mathbf{a}_i = (\Psi^\Phi)^T \mathbf{v}_i$ is an $(M \times m)$ -dimensional column vector and it is denoted by $\mathbf{a}_i = [\alpha_i^1, \alpha_i^2, \dots, \alpha_i^{M \times m}]^T$. Thus, the solutions $\mathbf{v}_i$ lie in the span of $\hat{\Phi}(\mathbf{A}_k^l)^T, k = 1, \dots, M; l = 1, \dots, m$. That is

$$\mathbf{v}_i = \sum_{k=1}^M \sum_{l=1}^m \alpha_i^{k \times l} \hat{\Phi}(\mathbf{A}_k^l)^T \quad (3.24)$$

Multiply $\hat{\Phi}(\mathbf{A}_g^h)^T$ on both sides of Equation (3.14), we get,

$$\lambda_i (\hat{\Phi}(\mathbf{A}_g^h)^T \cdot \mathbf{v}_i) = (\hat{\Phi}(\mathbf{A}_g^h)^T \cdot \mathbf{C}^\Phi \mathbf{v}_i) \quad (3.25)$$

That is,

$$\begin{aligned} & \lambda_i \sum_{k=1}^M \sum_{l=1}^m \alpha_i^{k \times l} (\hat{\Phi}(\mathbf{A}_g^h)^T \cdot \hat{\Phi}(\mathbf{A}_k^l)^T) \\ &= (\hat{\Phi}(\mathbf{A}_g^h)^T \cdot [\frac{1}{M} \sum_{t=1}^M \hat{\Phi}(\mathbf{A}_t)^T \hat{\Phi}(\mathbf{A}_t) \sum_{k=1}^M \sum_{l=1}^m \alpha_i^{k \times l} \hat{\Phi}(\mathbf{A}_k^l)^T]) \end{aligned}$$

$$\begin{aligned}
&= (\hat{\Phi}(\mathbf{A}_g^h)^T \cdot [\frac{1}{M} \sum_{p=1}^M \sum_{q=1}^m \hat{\Phi}(\mathbf{A}_p^q)^T \hat{\Phi}(\mathbf{A}_p^q) \sum_{k=1}^M \sum_{l=1}^m \alpha_i^{k \times l} \hat{\Phi}(\mathbf{A}_k^l)^T]) \\
&= \frac{1}{M} \sum_{k=1}^M \sum_{l=1}^m \alpha_i^{k \times l} (\hat{\Phi}(\mathbf{A}_g^h)^T \cdot \sum_{p=1}^M \sum_{q=1}^m \hat{\Phi}(\mathbf{A}_p^q)^T) (\hat{\Phi}(\mathbf{A}_p^q)^T \cdot \hat{\Phi}(\mathbf{A}_k^l)^T)
\end{aligned}$$

Defining an $(M \times m) \times (M \times m)$ matrix $\mathbf{K}$ by

$$\mathbf{K}_{(k \times l, p \times q)} = (\hat{\Phi}(\mathbf{A}_k^l)^T \cdot \hat{\Phi}(\mathbf{A}_p^q)^T) \quad (3.26)$$

The above equation can be converted into:

$$M\lambda_i \mathbf{K} \mathbf{a}_i = \mathbf{K}^2 \mathbf{a}_i \quad (3.27)$$

or

$$M\lambda_i \mathbf{a}_i = \mathbf{K} \mathbf{a}_i \quad (3.28)$$

Since $\mathbf{K}$ is positive semi-definite, $\mathbf{K}$'s eigenvalues will be nonnegative, the eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{M \times m}$ and the corresponding eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{M \times m}$ can be solved by diagonalizing $\mathbf{K}$, with λ_p being the first nonzero eigenvalue. We normalize $\mathbf{a}_p, \mathbf{a}_{p+1}, \dots, \mathbf{a}_{M \times m}$ by enforcing the unitilization of the corresponding $\mathbf{v}$ in $\mathbf{R}^f$, i.e., $(\mathbf{v}_d \cdot \mathbf{v}_d) = 1$ for all $d = p, \dots, M \times m$. In terms of $\mathbf{v}_i = \sum_{k=1}^M \sum_{l=1}^m \alpha_i^{k \times l} \hat{\Phi}(\mathbf{A}_k^l)^T$, this turns into: $1 = ((\sum_{k=1}^M \sum_{l=1}^m \alpha_d^{k \times l} \hat{\Phi}(\mathbf{A}_k^l)^T) \cdot (\sum_{p=1}^M \sum_{q=1}^m \alpha_d^{p \times q} \hat{\Phi}(\mathbf{A}_p^q)^T) = (\mathbf{a}_d \cdot \mathbf{K} \mathbf{a}_d) = \lambda_d (\mathbf{a}_d \cdot \mathbf{a}_d)$.

To extract the principal component of each row, we need to project each mapped row vector $\hat{\Phi}(\mathbf{A}_i^j)$ onto the eigenvectors $\mathbf{v}_k$ in $\mathbf{R}^f$, i.e., $(\mathbf{v}_k \cdot \hat{\Phi}(\mathbf{A}_i^j)) = \sum_{p=1}^M \sum_{q=1}^m \alpha_k^{p \times q} (\hat{\Phi}(\mathbf{A}_p^q)^T \cdot \hat{\Phi}(\mathbf{A}_i^j))$. Hence, similar to the essence of 2DPCA, K2DPCA performed on 2D images can be regarded as the clustered KPCA performed on the rows of all the training images. $\square$

After projecting each mapped row of the training and test images onto the first d eigenvectors in the feature space, an $m \times d$ feature matrix is obtained for each image. The nearest-neighborhood classifier is adopted for classification like B2DPCA.

3.6 Experimental Results and Discussion

3.6.1 Face Recognition on ORL and UMIST Face Databases

The proposed B2DPCA and K2DPCA methods are applied to the face recognition and are evaluated on the well-known ORL and UMIST face image databases. ORL face database contains images from 40 persons, each providing 10 different images. The pose, expression and facial details (e.g., with glasses or without glasses) variations are also included. The images are taken with a tolerance for some tilting and rotation of the face of up to 20 degrees. Moreover, there are also some variations in the scale of up to about 10 percent. Ten sample images of two persons from the ORL database are shown in Figure 3.1. UMIST face database consists of 564 images of 20 persons with large pose variations. In our experiment, 360 images with 18 samples for each subject are used to ensure that face appearance changes from profile to frontal orientation with a step of 5° separation (labelled from 1 to 18). The sample images for subject 1 are shown in Figure 3.2.

Figure 3.1: Ten sample images of two subjects in ORL database

Figure 3.2: Eighteen sample images of subject 1a from UMIST face database labelled by #1, #2,..., #18 from left to right

All images in ORL and UMIST databases are grayscale and normalized to a resolution of 56×46 pixels. The ORL database is employed to check whether the proposed methods have good generalization ability under the circumstances that the pose, expression, and face scale variations exist concurrently. The UMIST face database is used to examine the performance when face orientation varies significantly.

Table 3.2: Experiments on ORL database: the first row indicates the number of training sample for each person and the first column indicates the different methods used.

	1	2	3	4	5
PCA [12]	69.5%	82.5%	88.8%	92.1%	94.1%
KPCA [58]	69.5%	82.5%	88.8%	92.1%	94.2%
LDA [30]	——-	75.8%	87.0%	90.1%	91.7%
KLDA [59]	——-	85.5%	92.2%	95.6%	97.5%
2DPCA [9]	72.5%	84.5%	89.9%	93.1%	95%
KDDA [66]	——-	85.0%	88.6%	92.8%	96.0%
DCV [49]	——-	74.3%	82.9%	87.1%	88.7%
B2DPCA	72.8%	85.1%	90.3%	93.5%	95.4%
K2DPCA	74.5%	86.9%	92.0%	94.6%	96.2%

To test the recognition performance with different training numbers on *ORL*, k ($1 \leq k \leq 5$) images of each subject are randomly selected for training and the remaining $(10-k)$ images of each subject for testing. When $2 \leq k \leq 5$, 50 times of random selections are performed. When k equals 1, there are 10 possible selections for training. The final recognition rate is the average of all. The performance of the B2DPCA and K2DPCA compared with that of the state-of-the-art methods is listed in the Table 3.2, where the first row stands for the number of training samples for each subject and the first column stands for individual algorithm. LDA, KFDA, Kernel Direct Discriminant Analysis (KDDA) [66] and Discriminant Common Vector (DCV) [49] cannot be used when there is only one training sample for each subject.

Two experiments, with small number of training samples (2 and 3 for each subject), are conducted on UMIST database. When the number of training samples for each individual is 2, we select $\{\#5, \#14\}$ face images of each subject for training, the remaining for test. When the number of training samples is 3 for each subject, six groups are selected for training, i.e., $1\{\#1, \#7, \#13\}$, $2\{\#2, \#8, \#14\}$, $3\{\#3, \#9, \#15\}$, $4\{\#4, \#10, \#16\}$, $5\{\#5, \#11, \#17\}$ and $6\{\#6, \#12, \#18\}$, the remaining images corresponding to each group are used for test. The performance of the B2DPCA and K2DPCA is compared with that of the state-of-the-art methods in the Table 3.3, where the first row stands for the different training set and the first

Table 3.3: Experiments on UMIST database: the first row indicates the training images for each person and the first column indicates the different methods used.

	#5, #14	#1, #7, #13	#2, #8, #14	#3, #9, #15	#4, #10, #16	#5, #11, #17	#6, #12, #18
PCA [12]	80.3%	82.7%	89.7%	90.7%	90.7%	88.0%	86.0%
KPCA [58]	80.9%	86.0%	87.0%	91.0%	92.0%	89.3%	87.3%
LDA [30]	77.5%	90.0%	91.3%	95.0%	96.3%	94.3%	91.7%
KLDA [59]	92.5%	94.7%	96.7%	98.3%	99.0%	98.0%	97.3%
2DPCA [9]	90.3%	91.0%	93.0%	95.0%	95.0%	93.7%	92.3%
KDDA [66]	87.8%	94.0%	96.0%	95.7%	97.3%	95.7%	95.7%
DCV [49]	84.1%	89.7%	93.7%	97.7%	94.7%	92.7%	88.0%
B2DPCA	90.7%	91.7%	93.4%	95.3%	95.8%	94.0%	92.8%
K2DPCA	92.7%	94.0%	94.3%	95.7%	97.0%	95.7%	94.0%

column stands for the individual algorithm.

The Gaussian RBF kernel is adopted in the K2DPCA in all the experiments, and the optimal results are achieved when the width, δ , of the kernel is 2.718. The optimal dimensions of $\mathbf{U}_{opt}$ and $\mathbf{V}_{opt}$ of B2DPCA in both experiments are 56×5 and 56×5 . Therefore, the size of the extracted feature matrix for each image is 5×5 . For both experiments, the nearest-neighborhood classification criterion is adopted and the distance function between any two feature matrices is the same as the one used in 2DPCA. From the experiments, we found that KFDA is the best of all the compared algorithms when the number of training samples for each subject is over 2. B2DPCA is slightly better than 2DPCA in recognition accuracy (improvement of 0.5% to 1.0%), but it can achieve higher compression rate, which is important in recognition step since higher compression rate will result in faster comparison between the probe image and the gallery images. K2DPCA does outperform 2DPCA and KPCA as explained in Section 3.1. We also find that K2DPCA is superior to LDA and DCV. In addition, K2DPCA is comparable to KDDA in all the experiments that we have conducted. K2DPCA even outperforms KFDA when the number of training sample is 2. According to the recognition accuracy, K2DPCA is better than B2DPCA in generalization ability due to its nonlinear property.

3.6.2 Effect of d -value on Recognition Rate

We set a common d for both l and r in B2DPCA. Therefore, the final feature image obtained from B2DPCA for each image is a $d \times d$ square matrix. A large value of d will result in a small compression rate while a small value of d will lose some important information for classification. To illustrate this situation, a number of experiments are conducted on the two databases. The results are shown in Figure 3.3, where the x -axis denotes the d -value and the y -axis denotes the recognition rate. Three experiments with different number of training samples (2, 3 and 4 training images for each subject respectively) for each subject are carried out on *ORL* database. Three experiments with different training set (1{#1, #7, #13}, 3{#3, #9, #15}, 5{#5, #11, #17}) are conducted on *UMIST*. From Figure 3.3, when the d -value is about 5, the B2DPCA will achieve the highest recognition rate. When d is larger, the recognition rate is nearly constant. To ensure an efficient classification and high compression rate, d is set to be the value of 5.

Figure 3.3: The effect of different d -value on recognition rate of B2DPCA

3.6.3 Convergence Analysis

Figure 3.4: Convergence of B2DPCA

The mean reconstruction error of the images, $\mathcal{E}$, can be used as a measure to check the convergency of the B2DPCA algorithm. As the iteration proceeds, the $\mathbf{C}_v$, $\mathbf{C}_u$, their eigenvectors corresponding to their top eigenvalues, $\mathbf{V}_{Opt}$, $\mathbf{U}_{Opt}$ and the reconstruction error are updated timely. In this experiment, the reconstruction error is shown to be converging as the iteration proceeds. The mean reconstruction error is defined as $\mathcal{E} = \frac{1}{M} \sum_{i=1}^M \|\mathbf{A}_i - \mathbf{U}\mathbf{B}_i\mathbf{V}^T\|_F$. For simplicity, we set $d=10$ for all cases. The same six experiments as those of Section 3.6.2 are conducted and the results are reported in Figure 3.4, where the x -axis denotes the iteration number and the y -axis denotes the error. From Figure 3.4, it can be found that the error converges after two iterations. Therefore, B2DPCA can achieve fast feature extraction even the bilateral operation is performed.

Figure 3.5: First row: raw images. Second row and fourth row: image reconstructed and compressed by 2DPCA using 2 and 8 *principal component (vectors)* respectively. Third row and fifth row: image reconstructed and compressed by B2DPCA with $d = 10$ and $d = 20$ respectively.

3.6.4 Reconstruction and Compression of Face Images

2DPCA is an excellent dimension-reduction tool for image processing, compression, storage and transmission. In this section, we compare the compression rate and reconstruction effect of B2DPCA with that of 2DPCA. Figure 3.5 shows the reconstruction effect of them, where the raw images lie on the first row and the reconstructed image by 2DPCA using 2 and 8 *principal component (vectors)* are shown in the second and fourth rows respectively. The reconstructed images by B2DPCA with $d = 10$ and $d = 20$ are shown in the third and fifth rows. Therefore, the second and third rows have almost the same compression rate since $\frac{56 \times 46}{56 \times 2} \approx \frac{56 \times 46}{10 \times 10}$, while the fourth and fifth rows have almost the same compression rate since $\frac{56 \times 46}{56 \times 8} \approx \frac{56 \times 46}{20 \times 20}$. But the effect of reconstruction by B2DPCA on the third and fifth rows are much better than that by 2DPCA on the second and fourth rows respectively.

3.7 Concluding Remarks

By analyzing the drawbacks of the 2D Principal Component Analysis (2DPCA), a framework of Generalized 2D Principal Component Analysis (G2DPCA) is proposed to extend the original 2DPCA in three ways. Firstly, the essence of 2DPCA is clarified and it explains why 2DPCA is better than PCA in terms of the accuracy in recognition. Secondly, a bilateral 2DPCA scheme (B2DPCA) is introduced to remove the necessity of more coefficients in representing an image in 2DPCA than in PCA. Thirdly, a kernel 2DPCA scheme (K2DPCA) is introduced to remedy the shortage of 2DPCA in exploring the higher-order statistics among the rows/columns of the input data. Experimental results demonstrate the excellent performance of the G2DPCA algorithms.

Chapter 4

2D Fisher Discriminant Analysis

In this chapter, the 2D Fisher Discriminant Analysis is proposed for face recognition with small number of training samples. This chapter is organized as follows: The Small Sample Size Problem and the review on the previous solutions to SSS problem are given in Section 4.1. The Unilateral 2D Fisher Discriminant Analysis and Bilateral 2D Fisher Discriminant Analysis are given in Section 4.2 and Section 4.3. The essence of 2DFDA is revealed in Section 4.4. The Kernel 2D Fisher Discriminant Analysis is given in Section 4.5. Experiment results and discussion are listed in Section 4.6 and the concluding remarks are drawn in the final section.

4.1 Introduction

4.1.1 The Small Sample Size Problem

When only t samples are available in an n -dimensional vector space with $t < n$, the sample covariance matrix $\hat{\mathbf{C}}$ is calculated from the samples as

$$\hat{\mathbf{C}} = \frac{1}{t} \sum_{i=1}^t (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \quad (4.1)$$

where $\mathbf{m}$ is the mean of all the samples. $(\mathbf{x}_i - \mathbf{m})$'s are not linearly independent because they are related by $\sum_{i=1}^t (\mathbf{x}_i - \mathbf{m}) = 0$. That is, $\hat{\mathbf{C}}$ is a function of $(t - 1)$ or less linearly independent vectors. Therefore, the rank of $\hat{\mathbf{C}}$ is $(t - 1)$ or less. This problem, which is called a Small Sample Size (SSS) problem [56], is often encountered in visual pattern recognition area. In the Fisher linear discriminant analysis of visual patterns, the criterion of measuring the discriminatory power of the projection vectors is to maximize the between-class scatter and in the meantime to minimize the within-class scatter of the projected samples. However, the within-class scatter matrix, $\mathbf{S}_w$, is often singular because of the SSS problem, which introduces problems in solving the optimally discriminating directions.

4.1.2 Previous Solutions to the SSS problem

To solve the SSS problem in the LDA-based schemes, various schemes have been proposed so far. Swets and Weng's discriminant eigenfeatures [28], Belhumeur et al.'s Fisherface [30] and Zhao's discriminant component analysis [31] all used a two-stage PCA+LDA approach. Using PCA, the high-dimensional face data are projected to a low-dimensional space and then LDA can be applied to this PCA subspace. However, the removed subspace may contain some usefully discriminative information. In other words, this removal may result in a loss of discriminative information. Chen et al. [38] suggested that the null space spanned by the eigenvectors of $\mathbf{S}_w$ with zero eigenvalues contains the most discriminative information, and an LDA method in the null space of $\mathbf{S}_w$ was proposed, which is called N-LDA. However, as explained in [38], with the existence of noise, when the number of training samples is large, the null space of $\mathbf{S}_w$ becomes small, and so much discriminative information outside this null space will be lost. Another shortcoming of this approach is that it involves solving the eigenvalue problem for a very large matrix.

Yu et al. [39] proposed an algorithm called Direct-LDA (D-LDA) which also incorporates the concept of null space. It first removes the null space of the between-class

scatter matrix, $\mathbf{S}_b$, under the assumption that $\mathbf{S}_b$ contains no discriminative information, and seeks a projection to minimize $\mathbf{S}_w$ in the subspace of $\mathbf{S}_b$. As the rank of $\mathbf{S}_b$ is smaller than that of $\mathbf{S}_w$, removing the null space of $\mathbf{S}_b$ may lose partial or entire null space of $\mathbf{S}_w$, which is very likely to be full-rank after the removal. In addition, the assumption made by D-LDA that the projection vectors are restricted in the subspace spanned by those vectors connecting class centers cannot be generally satisfied [41].

Wang and Tang [41] presented a random sampling LDA method. Compared with the previous methods, it does not directly solve the problem of the singularity of $\mathbf{S}_w$. But it can be regarded as an enforcement of weak classifiers. This method concludes that both Fisherface and N-LDA encounter respective over-fitting problem. A random subspace method [42] and a random bagging approach are proposed to solve the over-fitting problem. A fusion rule is adopted to combine these two kinds of random sampling based classifiers.

A dual-space LDA approach [43] for face recognition was proposed to take full advantage of the discriminative information in the face space. Based on a probabilistic visual model, the eigenvalue spectrum in the null space of within-class scatter matrix is estimated, and the discriminant analysis is simultaneously applied in the principal and null subspaces of the within-class scatter matrix. The two sets of discriminative features are then combined for recognition purposes.

More recently, a novel scheme which is called Discriminant Common Vectors (DCV) method [49] was proposed to solve the SSS problem in face recognition. In this scheme, two algorithms are given to extract the discriminant common vectors for representing each person in the training set of the face database. One algorithm uses the within-class scatter matrix of the samples in the training set while the other uses the subspace method and the Gram-Schmidt orthogonalization procedure. The discriminative common vectors are used for classification of new faces. However, this scheme is essentially a null space based method. As stated in [49], the success of this method depends on the size of the null space of the within-class scatter matrix.

Since there will not be sufficient space for obtaining the optimal projection vectors when the size of the null space is small, recognition rates are expected to be poor.

The representations in the above linear subspaces methods are based on second-order statistics of the image set, and do not address higher-order statistical dependencies such as the relationships among three or more pixels. Since much of the important information may be contained in the high-order dependencies among pixels of an image, there has been a trend to investigate the nonlinear subspace approaches in recognition. Using integral operator kernel function, Scholkopf et al. [58] investigated the nonlinear component analysis by efficiently extending the classical PCA to Kernel Principle Component Analysis (KPCA) in a high-dimensional feature space, which is related by some nonlinear map. Yang [59] introduced the use of the Kernel Eigenface and Kernel Fisherface for learning low dimensional representations for face recognition. Specifically, the Kernel Fisherface method not only extracts high-order statistical dependencies of samples as features, but it also maximizes the class separation when these features are projected to a lower dimensional space. Lu et al. [66] gave a Kernel Direct Discriminant Analysis (KDDA) algorithm by extending the linear D-LDA method to a high-dimensional feature space.

4.1.3 2D Data Representation Model

A common property of the methods mentioned in Section 4.1.2 is that the vector-based data representation model is adopted. Under this model, each datum is represented as a vector and the collection of data is modelled as a single data matrix. The representation by vectors has explicitly physical meaning since it allows one to compute the similarity between two data according to the Euclidean distance or some other similarity metrics. However, it also brings the disadvantage that the feature dimension is very high in image classification and retrieval. Because of the Curse of Dimensionality and Small Sample Size problem, the performance degrades as the dimensionality increases and the sample size decreases.

Recently, a 2D Principal Component Analysis (2DPCA) [9] method has been proposed. 2DPCA is based on 2D image matrices rather than column vectors as opposed to the traditional PCA. By defining the covariance matrix in the way as in [9], the essence of 2DPCA performed on the image matrices can be viewed as a clustered PCA performed on all the rows of the training samples. The advantage of 2DPCA compared with PCA is that the covariance matrix is quite small and can be evaluated more accurately. This makes the 2DPCA achieve higher recognition rate than PCA. We have concluded the intrinsic reasons in [68] and [69]. Firstly, as the dimension of the row of each image is much smaller than that of the entire image vector, the dilemma of Curse of Dimensionality diminishes. Secondly, as the input feature vectors to be analyzed are factually the rows of all the training images, the feature set is significantly enlarged. Therefore, the effect by the SSS problem is reduced in 2DPCA. Thirdly, the 2D spatial information is kept complete by reserving the 2D image matrix rather than disintegrating it. However, like PCA, 2DPCA is only good at image representation rather than image discrimination. When there are large pose and illumination variations in face images, the top eigenvectors in 2DPCA do not model the identity information but these external variations and it can be expected that 2DPCA will be inferior to LDA. Our experiments will verify this conclusion in the later section of this chapter.

4.1.4 2D Fisher Discriminant Analysis

To overcome the shortcoming in 2DPCA and to solve the SSS problem in LDA based algorithms, a novel 2DFDA framework containing Unilateral 2DFDA (U2DFDA) and Bilateral 2DFDA (B2DFDA) is proposed in this chapter. Similar to 2DPCA, 2DFDA constructs the between-class and within-class scatter matrix using the image matrices. Differing from 2DPCA, the Fisher's criterion is adopted to find the discriminatively optimal information. In addition, if each row/column of the image is viewed as an new computational unit, it can be shown that the proposed 2DFDA performed on the 2D image matrices is essentially a clustered FDA performed on all

the rows/columns of the training images. In contrast to the $\mathbf{S}_b$ and $\mathbf{S}_w$ of FDA, the $\mathbf{S}_b$ and $\mathbf{S}_w$ in 2DFDA are not singular. As a result, the 2DFDA has three advantages over the 2DPCA and FDA based algorithms. Firstly, the features are extracted using Fisher discriminant analysis instead of PCA. Thus, the discriminating ability is better than 2DPCA. Secondly, it does not encounter the Curse of Dimensionality and the SSS problem is significantly reduced. Thirdly, it takes full advantage of the discriminative information in the face image space, and does not discard any subspace which may be useful for recognition purposes.

Within this framework, B2DFDA can achieve better performance than its unilateral counterpart because the incorporation of the bilateral projections will extract more discriminant information. Experiment results on ORL face database, Yale face database B and UMIST face database will clearly demonstrate these advantages. At the same time, Ye et al. [67] have also described a similar idea using a 2D-matrix based data representation method. Differing from our method, they adopt a scheme of simultaneously bilateral projection, and an iterative process is performed to solve the two optimal projection metrics. However, the simultaneous projection is not necessarily better than the asynchronous one adopted in this chapter. The reasons are as follows: Firstly, Ye's method needs an initialization which could produce the locally optimal features. Instead, our method does not require any initialization. Secondly, Ye's method may lose some discriminant information because the simultaneous projection scheme is essentially a re-projection of a body of discriminant features which will discard some discriminant information. Our experimental results show that the unilateral 2DFDA outperforms Ye's method.

4.2 Unilateral 2D Fisher Discriminant Analysis

Let $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d]$ denote an $m \times d$ projection matrix, where $\mathbf{w}_i$ is an m -dimensional unitary column vector, $i = 1, \dots, d$. The idea is to project image $\mathbf{X}$, an

$m \times n$ matrix, onto $\mathbf{W}$ by the following linear transformation:

$$\mathbf{Y} = \mathbf{W}^T \mathbf{X} \quad (4.2)$$

In doing so, we obtain a $d \times n$ projected feature matrix $\mathbf{Y}$ for each image. As in FDA, the discriminatory power of the projection vectors $\mathbf{W}$ can be measured by the Fisher criterion [56]. That is, maximizing the between-class scatter and in the meantime minimizing the within-class scatter of the projected samples. The Fisher criterion in 2DFDA as follows:

$$J(\mathbf{W}) = \frac{\det(\mathbf{PS}_b)}{\det(\mathbf{PS}_w)} \quad (4.3)$$

where $\mathbf{PS}_b$ and $\mathbf{PS}_w$ are the between-class and within-class covariance of the projected samples respectively, $\det(\cdot)$ denotes the determinant of a matrix.

It is also shown that the Fisher criterion can be written in another way.

Lemma 4.1: Let $\mathbf{S}_b$ and $\mathbf{S}_w$ be the between-class and within-class covariance of the original image matrix respectively. Then, $J(\mathbf{W}) = \frac{\det(\mathbf{W}^T \mathbf{S}_b \mathbf{W})}{\det(\mathbf{W}^T \mathbf{S}_w \mathbf{W})}$.

Proof: Let $\bar{\mathbf{M}}$ be the mean of all the training samples, $\bar{\mathbf{M}}_i$ be the mean of the i -th class, $\bar{\mathbf{M}}^p$ be the mean of all the projected samples, $\bar{\mathbf{M}}_i^p$ be the mean of the i -th projected class. Then, $\mathbf{PS}_b = \sum_{i=1}^L L_i (\bar{\mathbf{M}}_i^p - \bar{\mathbf{M}}^p) (\bar{\mathbf{M}}_i^p - \bar{\mathbf{M}}^p)^T = \sum_{i=1}^L L_i (\mathbf{W}^T \bar{\mathbf{M}}_i - \mathbf{W}^T \bar{\mathbf{M}}) (\mathbf{W}^T \bar{\mathbf{M}}_i - \mathbf{W}^T \bar{\mathbf{M}})^T = \sum_{i=1}^L L_i \mathbf{W}^T (\bar{\mathbf{M}}_i - \bar{\mathbf{M}}) (\bar{\mathbf{M}}_i - \bar{\mathbf{M}})^T \mathbf{W} = \mathbf{W}^T \mathbf{S}_b \mathbf{W}$, where L is the total class number, L_i is the number of training samples in the i -th class, $\mathbf{S}_b = \sum_{i=1}^L L_i (\bar{\mathbf{M}}_i - \bar{\mathbf{M}}) (\bar{\mathbf{M}}_i - \bar{\mathbf{M}})^T$. Similarly, we have $\mathbf{PS}_w = \mathbf{W}^T \mathbf{S}_w \mathbf{W}$, where $\mathbf{S}_w = \sum_{i=1}^L \sum_{j=1}^{L_i} (\mathbf{X}_i^j - \bar{\mathbf{M}}_i) (\mathbf{X}_i^j - \bar{\mathbf{M}}_i)^T$. $\square$

Therefore, the Fisher criterion in Equation (4.3) can be converted to:

$$J(\mathbf{W}) = \frac{\det(\mathbf{W}^T \mathbf{S}_b \mathbf{W})}{\det(\mathbf{W}^T \mathbf{S}_w \mathbf{W})} \quad (4.4)$$

The vectors in $\mathbf{W}$ that maximize Equation (4.4) are called the *Optimal Discriminant*

Projection Axes.

Since the projection presented in Equation (4.2) is a unilaterally left-multiplication, the 2DFDA obtained in this way is called Unilateral Two Dimensional Fisher Discriminant Analysis (U2DFDA). For U2DFDA, we have the following theorem.

Lemma 4.2: $\mathbf{S}_w$ in U2DFDA is not singular.

Proof: $\mathbf{S}_w = \sum_{i=1}^L \sum_{j=1}^{L_i} (\mathbf{X}_i^j - \bar{\mathbf{M}}_i) (\mathbf{X}_i^j - \bar{\mathbf{M}}_i)^T$. It can also be written as $\mathbf{S}_w = \Phi_{S_w} \Phi_{S_w}^T$, where $\Phi_{S_w} = [\phi_1^{S_w}, \phi_2^{S_w}, \dots, \phi_L^{S_w}]$, $\phi_i^{S_w} = [(\mathbf{X}_i^1 - \bar{\mathbf{M}}_i), (\mathbf{X}_i^2 - \bar{\mathbf{M}}_i), \dots, (\mathbf{X}_i^{L_i} - \bar{\mathbf{M}}_i)]$, $i = 1, \dots, L$, $j = 1, \dots, L_i$ and L_i is the number of training samples in the i -th class. $\mathbf{X}_i^j$ is the j -th training sample in the i -th class. The dimension of Φ_{S_w} is $m \times (n \sum_{i=1}^L L_i)$, where m and n are the image's height and width. Since $\text{rank}(\Phi_{S_w}^T \Phi_{S_w}) = \text{rank}(\Phi_{S_w} \Phi_{S_w}^T) = \text{rank}(\Phi_{S_w}^T) = \text{rank}(\Phi_{S_w})$ and $\text{rank}(\Phi_{S_w}) = m$. It is known that $\text{rank}(\Phi_{S_w}) \leq \min(m, n \sum_{i=1}^L L_i)$ and $m \ll (n \sum_{i=1}^L L_i)$ in the area of visual pattern recognition, there is, $\text{rank}(\Phi_{S_w}) \leq m$. Further, we make an assumption that the rows of Φ_{S_w} are independent of each other (the experiment results will demonstrate that this assumption can be well satisfied in the benchmark databases). In addition, the dimension of $\mathbf{S}_w$ is $m \times m$, we can conclude that $\mathbf{S}_w$ is of full rank. $\square$

Theorem 4.1: The *Optimal Discriminant Projection Axis* in $\mathbf{W}_{opt}$ can be obtained by directly solving the following generalized eigen-value problem.

$$\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{W}_{opt} = \Lambda \mathbf{W}_{opt} \quad (4.5)$$

where Λ is the diagonal matrix whose diagonal elements are eigenvalues of $\mathbf{S}_w^{-1} \mathbf{S}_b$.

Proof: From Lemmas 4.1 and 4.2, it can be easily obtained. $\square$

The *Optimal Discriminant Projection Axis* of U2DFDA, $\mathbf{w}_1, \dots, \mathbf{w}_d$, can be used for feature extraction. For a given image sample $\mathbf{X}$, $\mathbf{Y} = \mathbf{W}^T \mathbf{X}$ is performed. $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_d]$, which is called *Fisher feature matrix* and $\mathbf{y}_i, i = 1, \dots, d$, is the i -th *Fisher feature vector*. In our framework, the nearest-neighbor classifier is used

for classification. Here, the distance between two arbitrary *Fisher feature matrices*, $\mathbf{Y}_i = [\mathbf{y}_1^i, \dots, \mathbf{y}_d^i]$ and $\mathbf{Y}_j = [\mathbf{y}_1^j, \dots, \mathbf{y}_d^j]$, is defined by

$$\text{dist}(\mathbf{Y}_i, \mathbf{Y}_j) = \sum_{k=1}^d \|\mathbf{y}_k^i - \mathbf{y}_k^j\|_2 \quad (4.6)$$

where $\|\mathbf{y}_k^i - \mathbf{y}_k^j\|_2$ denotes the Euclidean distance between $\mathbf{y}_k^i$ and $\mathbf{y}_k^j$.

Suppose that the *Fisher feature matrices* of the training samples are $\mathbf{Y}_1, \dots, \mathbf{Y}_M$, and that each of these samples is assigned a given identity (class) C_k . Given a test sample $\mathbf{Y}$, if $\text{dist}(\mathbf{Y}, \mathbf{Y}_l) = \min_{i=1}^M \text{dist}(\mathbf{Y}, \mathbf{Y}_i)$, and $\mathbf{Y}_l \in C_l$, then the resulting decision is $\mathbf{Y} \in C_l$.

In FDA based methods, the dimension for classification is fixed to $(C - 1)$, where C is the number of classes. However, the optimal number of *Fisher feature vector* in 2DFDA, d , is not fixed. Since the $\mathbf{S}_w$ is invertible, d can be at most equal to the image's height. However, the optimal d for classification is dependent on the database. In our experiments, we will discuss the optimal dimensions for different databases.

4.3 Bilateral 2D Fisher Discriminant Analysis

Section 4.2 describes a method of extracting the optimal discriminant directions via a left-multiplying U2DFDA. This section presents a right-multiplying operation. That is,

$$\mathbf{Y} = \mathbf{XW} \quad (4.7)$$

In fact, the right-multiplying U2DFDA can be converted into left-multiplying U2DFDA by transposing the image matrix. We explain this convertibility by using a single n -dimensional column vector $\mathbf{w}$. It is easy to prove that if $\mathbf{y}_r = \mathbf{Xw}$ and $\mathbf{y}_l = \mathbf{w}^T \mathbf{X}^T$, then $\mathbf{y}_l = \mathbf{y}_r^T$. Therefore, the right-multiplying *Fisher feature matrix*, $\mathbf{Y}_r = \mathbf{XW}_r$,

where $\mathbf{W}_r$ can be obtained by the following Fisher criterion:

$$\mathbf{W}_r = \arg \max \frac{\det(\mathbf{W}^T \mathbf{S}_b \mathbf{W})}{\det(\mathbf{W}' \mathbf{S}_w \mathbf{W})} \quad (4.8)$$

where $\mathbf{S}_b = \sum_{i=1}^L L_i (\bar{\mathbf{M}}_i - \bar{\mathbf{M}})^T (\bar{\mathbf{M}}_i - \bar{\mathbf{M}})$, $\mathbf{S}_w = \sum_{i=1}^L \sum_{j=1}^{L_i} (\mathbf{X}_i^j - \bar{\mathbf{M}}_i)^T (\mathbf{X}_i^j - \bar{\mathbf{M}}_i)$.

The computational methods for $\mathbf{S}_b$ and $\mathbf{S}_w$ are different in the left-multiplying and right-multiplying U2DFDA. For the left-multiplying U2DFDA, $\mathbf{S}_b = \sum_{i=1}^L L_i (\bar{\mathbf{M}}_i - \bar{\mathbf{M}}) (\bar{\mathbf{M}}_i - \bar{\mathbf{M}})^T$, $\mathbf{S}_w = \sum_{i=1}^L \sum_{j=1}^{L_i} (\mathbf{X}_i^j - \bar{\mathbf{M}}_i) (\mathbf{X}_i^j - \bar{\mathbf{M}}_i)^T$. For the right-multiplying U2DFDA, $\mathbf{S}_b = \sum_{i=1}^L L_i (\bar{\mathbf{M}}_i - \bar{\mathbf{M}})^T (\bar{\mathbf{M}}_i - \bar{\mathbf{M}})$, $\mathbf{S}_w = \sum_{i=1}^L \sum_{j=1}^{L_i} (\mathbf{X}_i^j - \bar{\mathbf{M}}_i)^T (\mathbf{X}_i^j - \bar{\mathbf{M}}_i)$. However, we may find that either in left-multiplying U2DFDA or right-multiplying U2DFDA, the computations of $\mathbf{S}_b$ and $\mathbf{S}_w$ solely emphasize the dependency (correlation) among either the row or the column vectors of the image matrix while neglect the other one. Therefore, it may lose some discriminant information. Considering this, a bilateral-projection scheme, called Bilateral 2D Fisher Discriminant Analysis (B2DFDA), is proposed to remedy the drawback of U2DFDA. We have,

$$\begin{cases} \mathbf{Y}_l = \mathbf{W}_l^T \mathbf{X} \\ \mathbf{Y}_r = \mathbf{X} \mathbf{W}_r \end{cases} \quad (4.9)$$

where $\mathbf{W}_l = [\mathbf{w}_1^l, \mathbf{w}_2^l, \dots, \mathbf{w}_{d_l}^l]$, $\mathbf{W}_r = [\mathbf{w}_1^r, \mathbf{w}_2^r, \dots, \mathbf{w}_{d_r}^r]$ are the left- and right-multiplying optimal projection vectors respectively, d_l is the number of left-multiplying projection directions and it can be equal to the image's height at most. d_r is the number of right-multiplying projection directions and it equals to the image's width at most.

After the left- and right-multiplying U2DFDA, $\mathbf{Y}_l$ and $\mathbf{Y}_r$ are obtained for each image. They are combined for recognition purposes. The steps for recognition are listed in Table 4.1.

Table 4.1: Recognition using 2DFDA

Step 1. Transform $\mathbf{Y}_l$ and $\mathbf{Y}_r$ for each image into two 1D feature vectors. Thus, we get two groups of 1D feature vectors for all the images.
Step 2. Apply PCA transformation to the two groups of 1D feature vectors and we can obtain two lower-dimensional 1D vectors for each image (It is not a compulsory step in 2DFDA, however, to reduce the computational complexity, we adopt the PCA step).
Step 3. For each image, combine the two shorter 1D vector into a longer 1D vector. Recognition is carried out using nearest-neighbor classifier on the finally combined 1D vector.

4.4 The Essence of 2D Fisher Discriminant Analysis

We notice that the covariance matrices in U2DFDA appears to be physically meaningful in the matrix space rather than in the vector space. However, Theorem 4.2 will give another perspective to make the U2DFDA physically meaningful even in vector space, and will explicitly explain its essence.

Theorem 4.2: In the left-multiplying U2DFDA, the U2DFDA performed on the image matrices is essentially the clustered LDA method performed on the columns of the image matrices if each column is viewed as a computational unit.

Proof: Since $\mathbf{S}_w = \sum_{i=1}^L \sum_{j=1}^{L_i} (\mathbf{X}_i^j - \bar{\mathbf{M}}_i) (\mathbf{X}_i^j - \bar{\mathbf{M}}_i)^T$, or it can be written as $\mathbf{S}_w = \Phi_{S_w} \Phi_{S_w}^T$. $\Phi_{S_w} = [\phi_1^{S_w}, \phi_2^{S_w}, \dots, \phi_L^{S_w}]$, and $\phi_i^{S_w} = [(\mathbf{X}_i^1 - \bar{\mathbf{M}}_i), (\mathbf{X}_i^2 - \bar{\mathbf{M}}_i), \dots, (\mathbf{X}_i^{L_i} - \bar{\mathbf{M}}_i)] = [((\mathbf{X}_i^1(:, 1) - \bar{\mathbf{M}}_i(:, 1))), \dots, ((\mathbf{X}_i^1(:, n) - \bar{\mathbf{M}}_i(:, n))), \dots, ((\mathbf{X}_i^{L_i}(:, 1) - \bar{\mathbf{M}}_i(:, 1))), \dots, (\mathbf{X}_i^{L_i}(:, n) - \bar{\mathbf{M}}_i(:, n)))]$, where $\mathbf{X}_i^j$ is the j -th training sample in the i -th class, and $\mathbf{A}(:, i)$ is the i -th column of matrix $\mathbf{A}$. Therefore, $\mathbf{S}_w$ is constructed directly by the columns of the zero-centered training image matrices. Similarly, $\mathbf{S}_b$ is also constructed using the columns.

Therefore, the left-multiplying U2DFDA performed on the image matrices can be viewed as the clustered LDA performed on the columns of all the training samples

if each column is viewed as a computational unit. $\square$

Similarly, the essence of right-multiplying U2DFDA is as follows.

Theorem 4.3: In the right-multiplying U2DFDA, the U2DFDA performed on the image matrices is essentially the clustered FDA method performed on the rows of the image matrices if each row is viewed as a computational unit.

Proof: The proof of Theorem 4.3 is similar to that of Theorem 4.2. $\square$

4.5 Kernel 2D Fisher Discriminant Analysis

Inspired by the essence of 2DFDA, it is desirable to keep the integral 2D structure of the image matrix rather than to disintegrate and then transform it to a long vector. Therefore, we adopt the scheme that only the rows/columns are mapped. That is, to kernelize LU2DFDA, each column of all the images is mapped onto a high-dimensional feature space; to kernelize RU2DFDA, the nonlinear mapping is performed on each row of all the images.

Let $\Phi : \mathbf{R}^t \rightarrow \mathbf{R}^f, f > t$ be a nonlinearly mapping, where t is the length of the row/column of an image and f can be arbitrarily large. The dot product in the feature space of $\mathbf{R}^f$ can be conveniently calculated via a pre-defined kernel function, such as the commonly used Gaussian RBF kernel function.

Without loss of generality, we take the kernelization of the LU2DFDA as an example. The kernelization of RU2DFDA is similar to that of the LU2DFDA. It is assumed that all the mapped data are zero-centered.

Theorem 4.4: Let $\Phi(\mathbf{X}_i^j)$ be the j -th mapped image of the i -th class. $\Phi(\mathbf{X}_i^j)(:, k)$ be the k -th column vector of it. $\Phi(\overline{\mathbf{M}})$ is the mean of all the training samples in the kernel space. $\Phi(\overline{\mathbf{M}}_i)$ is the mean of i -th class in the kernel space. The defined kernelized 2DFDA on the images is essentially the clustered KFDA performed locally on the columns of all the training image matrices.

Proof: The within-class covariance matrix, $\mathbf{S}_w^\Phi$ and the between-class covariance matrix, $\mathbf{S}_b^\Phi$, in $\mathbf{R}^f$ are:

$$\mathbf{S}_w^\Phi = \sum_{i=1}^L \sum_{j=1}^{L_i} (\Phi(\mathbf{X}_i^j) - \Phi(\overline{\mathbf{M}}_i))(\Phi(\mathbf{X}_i^j) - \Phi(\overline{\mathbf{M}}_i))^T \quad (4.10)$$

$$\mathbf{S}_b^\Phi = \sum_{i=1}^L L_i (\Phi(\overline{\mathbf{M}}_i) - \Phi(\overline{\mathbf{M}}))(\Phi(\overline{\mathbf{M}}_i) - \Phi(\overline{\mathbf{M}}))^T \quad (4.11)$$

To perform Fisher discriminant analysis in $\mathbf{R}^f$, it is equivalent to maximizing:

$$\mathbf{J}(\vec{\omega}) = \frac{\vec{\omega}^T \mathbf{S}_b^\Phi \vec{\omega}}{\vec{\omega}^T \mathbf{S}_w^\Phi \vec{\omega}} \quad (4.12)$$

Because any solution $\vec{\omega} \in \mathbf{R}^f$ must lie in the span of the columns of all the training samples, i.e., there exist coefficients $\vec{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_{(n \sum_{i=1}^L L_i)}]^T$ such that

$$\begin{aligned} \vec{\omega} = & [[\Phi(\mathbf{X}_1^1(:, 1)), \Phi(\mathbf{X}_1^1(:, 2)), \dots, \Phi(\mathbf{X}_1^1(:, n))], \\ & , \dots, [\Phi(\mathbf{X}_L^{L_i}(:, 1)), \Phi(\mathbf{X}_L^{L_i}(:, 2)), \dots, \Phi(\mathbf{X}_L^{L_i}(:, n))]] \vec{\alpha} \end{aligned} \quad (4.13)$$

Therefore, the projection of $\Phi(\overline{\mathbf{M}}_i(:, k))$, the k -th column of the i -th class mean, onto $\vec{\omega}$, i.e., $\vec{\omega}^T \Phi(\overline{\mathbf{M}}_i(:, k))$, can be written as:

$$\vec{\alpha}^T \begin{bmatrix} \Phi(\mathbf{X}_1^1(:, 1))^T \\ \Phi(\mathbf{X}_1^1(:, 2))^T \\ \dots \\ \Phi(\mathbf{X}_L^{L_i}(:, n))^T \end{bmatrix} \frac{1}{L_i} \sum_{j=1}^{L_i} \Phi(\overline{\mathbf{M}}_i^j(:, k)) = \vec{\alpha}^T \mathcal{M}_i^k \quad (4.14)$$

and the projection of $\Phi(\overline{\mathbf{M}}(:, k))$, the k -th column of the total class mean, onto $\vec{\omega}$,

i.e., $\vec{\omega}^T \Phi(\bar{\mathbf{M}}(:, k))$ can be written as

$$\vec{\alpha}^T \begin{bmatrix} \Phi(\mathbf{X}_1^1(:, 1))^T \\ \Phi(\mathbf{X}_1^1(:, 2))^T \\ \dots \\ \Phi(\mathbf{X}_L^{L_i}(:, n))^T \end{bmatrix} \frac{1}{\sum_{i=1}^L L_i} \sum_{i=1}^L \sum_{j=1}^{L_i} \Phi(\bar{\mathbf{M}}_i^j(:, k)) = \vec{\alpha}^T \mathcal{M}^k \quad (4.15)$$

Thus, the numerator of Equation (4.12), $\vec{\omega}^T \mathbf{S}_b^\Phi \vec{\omega}$, can be converted into:

$$\vec{\omega}^T \left(\sum_{i=1}^L L_i (\Phi(\bar{\mathbf{M}}_i) - \Phi(\bar{\mathbf{M}})) (\Phi(\bar{\mathbf{M}}_i) - \Phi(\bar{\mathbf{M}}))^T \right) \vec{\omega} = \vec{\omega}^T \mathbf{Q} \mathbf{Q}^T \vec{\omega} \quad (4.16)$$

where

$$\mathbf{Q} = [(\sqrt{L_1}(\Phi(\bar{\mathbf{M}}_1) - \Phi(\bar{\mathbf{M}}))), \dots, (\sqrt{L_L}(\Phi(\bar{\mathbf{M}}_L) - \Phi(\bar{\mathbf{M}})))] \quad (4.17)$$

or it can be written in another form,

$$\vec{\omega}^T \mathbf{Q} \mathbf{Q}^T \vec{\omega} = \vec{\alpha}^T \mathbf{K}_b \vec{\alpha} \quad (4.18)$$

where $\mathbf{K}_b = \sum_{i=1}^L L_i (\mathcal{M}_i - \mathcal{M})(\mathcal{M}_i - \mathcal{M})^T$ and $\mathbf{K}_b$ is an $(n \sum_{i=1}^L L_i) \times (n \sum_{i=1}^L L_i)$ matrix, $\mathcal{M}_i = [\mathcal{M}_i^1, \dots, \mathcal{M}_i^n]$ and $\mathcal{M}_i^j = [\Phi(\mathbf{X}_1^1(:, 1)), \Phi(\mathbf{X}_1^1(:, 2)), \dots, \Phi(\mathbf{X}_L^{L_i}(:, n))]^T \Phi(\bar{\mathbf{M}}_i(:, j))$.

Similarly, the denominator of Equation (4.12),

$$\vec{\omega}^T \mathbf{S}_w^\Phi \vec{\omega} = \vec{\alpha}^T \mathbf{K}_w \vec{\alpha} \quad (4.19)$$

where $\mathbf{K}_w = \sum_{i=1}^L \sum_{j=1}^{L_i} (\mathcal{X}_i^j - \mathcal{M}_i)(\mathcal{X}_i^j - \mathcal{M}_i)^T$, $\mathcal{X}_i^j = [\mathcal{X}_i^j(:, 1), \dots, \mathcal{X}_i^j(:, n)]$, and $\mathcal{X}_i^j(:, k) = [\Phi(\mathbf{X}_1^1(:, 1)), \Phi(\mathbf{X}_1^1(:, 2)), \dots, \Phi(\mathbf{X}_L^{L_i}(:, n))]^T \Phi(\mathbf{X}_i^j(:, k))$

Thus, the maximization of Equation (4.12) is converted into the maximization of the following equation.

$$\mathbf{J}(\vec{\alpha}) = \frac{\vec{\alpha}^T \mathbf{K}_b^\Phi \vec{\alpha}}{\vec{\alpha}^T \mathbf{K}_w^\Phi \vec{\alpha}} \quad (4.20)$$

Similar to FDA, this problem can be solved by finding the leading eigenvectors of $\mathbf{K}_w^{-1}\mathbf{K}_b$ if the $\mathbf{K}_w$ is not singular. We can find from the above derivation of the newly defined K2DFDA that the K2DFDA performed on 2D image matrices is essentially the clustered KFDA method performed on the columns of all the images if each column is viewed as a point in the vector space. As the derivation above is based on the LU2DFDA, it is called KLU2DFDA. Similarly, the K2DFDA derived based on the RU2DFDA is called KRU2DFDA. We can also found that the KRU2DFDA is essentially the clustered KFDA method performed on the rows of all the images if each row is viewed as a point in the vector space. $\square$

However, because of the intrinsic shortcoming in constructing the $\mathbf{K}_w$, $\mathbf{K}_w$ is singular. To solve this problem, $\mathbf{K}_w$ is replaced by $\mathbf{K}_w + \lambda\mathbf{I}$, where λ is a very small number and $\mathbf{I}$ is the identity matrix.

Similar to B2DFDA, a left-projection and a right-projection feature matrices can be obtained in K2DFDA. For KLU2DFDA, we project each column of the images to get the left-projection discriminant feature matrix for each image. For KRU2DFDA, we project each row of the images to get the right-projection discriminant feature matrix for each matrix. Therefore, a Kernel B2DFDA (KB2DFDA) can be used further by combining the KLU2DFDA and KRU2DFDA. In classification, the two feature matrices are combined for recognition via the nearest-neighborhood classifier as in B2DFDA.

4.6 Experimental Results and Discussion

The proposed U2DFDA, B2DFDA and KB2DFDA methods are tested on three commonly used face image databases (ORL [155], Yale Face Database B [158] and UMIST face database [159]). The ORL database is used to evaluate the performance of 2DFDA under conditions where the pose, expression and scale vary. The Yale Face Database B is used to examine the performance of 2DFDA when illumination

varies significantly. Finally the multi-view UMIST face database is used to judge the performance of 2DFDA when there are large pose variations in face images.

The ORL face database contains images from 40 individuals, 10 images for each subject. For some subjects, the images are taken at different times. The facial expressions and facial details (with/without glasses) also vary. The images are taken with a tolerance for some tilting and rotation of up to 20 degrees. Moreover, there is also some variation in the scale of up to about 10 percent. All images are grayscale and normalized to a resolution of 56×46 pixels. Ten sample images of two persons from the ORL database are shown in Figure 3.1.

Yale Face Database B contains 5760 images of 10 subjects each seen under 576 viewing conditions (9 poses $\times$ 64 illumination conditions). Twenty sample images of two persons from the Yale Face Database B are shown in Figure 4.1. In our experiment, altogether 640 images for 10 subjects are used (64 illumination conditions under the same frontal pose). All the images are preprocessed and normalized by translation, rotation, and scaling such that the two outer eye corners are in fixed positions. The normalized image size is 50×40 .

Figure 4.1: Twenty sample images of two subjects in Yale face database B

The UMIST Face database consists of 564 images of 20 persons with large pose variations. In our experiment, only 360 images with 18 samples for each subject are used so that face appearance changes from profile to frontal orientation (with label from #1 to #18). The sample images for subject 1 are shown in Figure 3.2

Table 4.2: Five Experiments on ORL and YaleB

Experiment list	Description
E1.	Investigate the performance of U2DFDA with different number of <i>Fisher feature vectors</i> (Figure 4.2).
E2.	Compare B2DFDA with U2DFDA (Figure 4.2).
E3.	Compare 2DFDA with the state-of-the-art subspace methods with different number of training samples for each subject (Tables 4.3 and 4.4. Figures 4.3, 4.4, 4.5 and 4.6).
E4.	Compare the optimal results from 2DFDA and the optimal ones from FDA (Figures 4.7, 4.8 and 4.9).
E5.	Compare B2DFDA and U2DFDA with Ye’s 2DLDA [67] (Figures 4.10 and 4.11).

4.6.1 Experiments on ORL and YaleB face databases

Random Grouping of Training and Testing Sets

To test the recognition performance with different training numbers, k ($2 \leq k \leq 9$ for ORL database and $2 \leq k \leq 12$ for Yale face database B) images of each subject are randomly selected for training and the remaining $(p-k)$ (p is the total number of images for each subject, for ORL database, p equals 10; for Yale face database B, p is 64) images of each subject for testing. For each number k , 50 iterations of random selections are performed on ORL database and 100 iterations for Yale face database B. The final recognition rate is the average of all.

Five experiments on ORL and YaleB Respectively

We have designed five experiments to evaluate the performance of 2DFDA on ORL and YaleB face database. They are listed in Table 4.2.

In the first experiment, without loss of generality, the right-multiplying mode is used in U2DFDA. For ORL database, the maximum size of the *Fisher feature matrix* is 56×46 , i.e., containing at most 46 56-dimensional *Fisher feature vectors*. For Yale

face database B, the maximum size of the *Fisher feature matrix* is 50×40 , i.e., containing at most 40 50-dimensional *Fisher feature vectors*. We change the number of *Fisher feature vectors* from 1 to 46 for ORL database and from 1 to 40 for Yale face database B to see the effect on performance. It has been shown via theoretical analysis that there is no SSS problem in 2DFDA. In order to verify this argument again through experiment, we focus on testing the performance of 2DFDA when there are only few training samples for each subject, e.g. only 2, 3 or 4 training samples for each subject. Figure 4.2 (a) to (c) show the performance of U2DFDA on ORL database. The optimal number of the *Fisher feature vectors* in all the three trials is found to be 3. Figure 4.2 (d) to (f) show the performance of U2DFDA on Yale face database B. The optimal number of the *Fisher feature vectors* in the three trials is found to be 31, 27 and 23 respectively. From the experimental results, it can be seen that the optimal number of *Fisher feature vectors* in U2DFDA is different for different database. Even on the same database, the optimal number will vary according to the number of the training samples for each subject.

The second experiment compares B2DFDA with U2DFDA. By fixing the optimal number of the *Fisher feature vectors* of the right-multiplying U2DFDA, the number of the *Fisher feature vectors* of the left-multiplying U2DFDA is changed from 1 to 56 for ORL database and from 1 to 50 for Yale face database B. Then B2DFDA is applied. Figure 4.2 (a) to (c) show the comparison of B2DFDA and U2DFDA on ORL database while Figure 4.2 (d) to (f) show the comparison results on Yale face database B. From these experiments, it can be found that B2DFDA achieves higher recognition rate than U2DFDA, e.g., with an increase of up to 5 percent on Yale face database B.

The aim of the third experiment is to test the performance of U2DFDA, B2DFDA and KB2DFDA with respect to different number of training samples for each subject. A comparison is made between 2DFDA and the state-of-the-art linear and kernel subspace methods. In Fisherface [30], the size of PCA subspace is constrained to $(N - C)$, the dimension for LDA is set to be $(C - 1)$, where N is the total number

Figure 4.2: Comparison between U2DFDA and B2DFDA with different number of *Fisher feature vectors*. (a)-(c): Two, three and four training samples respectively for each subject on ORL database; (d)-(f): Two, three and four training samples respectively for each subject on Yale face database B.

of the training samples, C is the number of the classes. In D-LDA [39] and N-LDA [40], the dimensions for LDA are both set to be $(C - 1)$. In 2DPCA and U2DFDA, we managed to find the optimal numbers of *Eigen feature vector* and *Fisher feature vector* which give the best classification. The optimal numbers of 2DPCA and U2DFDA for ORL database and Yale face database B are reported in Figure 4.3 and Figure 4.4 respectively. In B2DFDA, the numbers of the right- and left-multiplying *Fisher feature vector* are set to be equal, both being the optimal number in U2DFDA. Two observations can be noticed from Figure 4.3 and Figure 4.4. Firstly, on the ORL database, with the increase in the number of training samples for each subject, the optimal number also increases in 2DPCA. On the contrary, the optimal number decreases in U2DFDA. On the Yale face database B, the optimal number equals to the maximum number of Eigen feature vectors in 2DPCA. However, the optimal number drops as the number of training samples increases in U2DFDA. Secondly, the optimal number of U2DFDA is always smaller than that of 2DPCA. Hence, from the point view of storage-space requirement and classification time, U2DFDA is more efficient than 2DPCA.

Figure 4.5 shows the average recognition rate on the ORL database. It can be seen that the performance of U2DFDA and B2DFDA is much better than the other linear methods, especially when the number of training sample is small. 2DFDA outperforms the other LDA based algorithms by up to 7 to 12 percent, and surpasses the 2DPCA by up to 3.5 percent. We can also find that B2DFDA does outperform U2DFDA as expected by the theoretical analysis. 2DPCA is superior to the LDA based algorithms on this database where there are no significant illumination variations. Figure 4.6 shows the average recognition rate on Yale face database B. B2DFDA is much better than U2DFDA, with an increase of 2 to 5 percent.

From Figure 4.5 and 4.6, we may find that the performance of 2DFDA is the best among all the compared linear subspace methods when the number of training sample for each subject is small. However, Fisherface will outperform B2DFDA when the number of training samples for each subject increases to a larger one (e.g., 12

Figure 4.3: Optimal number of *Fisher feature vector* and *Eigen feature vector* in U2DFDA and 2DPCA with different number of training samples for each subject on the ORL database

Figure 4.4: Optimal number of *Fisher feature vector* and *Eigen feature vector* in U2DFDA and 2DPCA with different number of training samples for each subject on the Yale face database B

Figure 4.5: Recognition rate of U2DFDA and B2DFDA compared with the other subspace methods on the ORL face database

for YaleB face database). This is because, in Fisherface method, the null space of $\mathbf{S}_w$ is discarded, whose rank is $(C - 1)$. When the number of training samples for each subject is small, the rank of null space of $\mathbf{S}_w$ is comparable to the rank of range space of $\mathbf{S}_w$. Therefore, discarding the whole null space of $\mathbf{S}_w$ is equivalent to losing a large quantity of discriminant information. However, with the number of training samples increasing, the rank of null space of $\mathbf{S}_w$ is much smaller than the rank of range space of $\mathbf{S}_w$ and discarding the whole null space of $\mathbf{S}_w$ will lose relatively little useful information.

We also find the performance of 2DPCA on Yale face database B is not as good as that on ORL database. This is consistent with our previous analysis that 2DPCA is a good pattern representation method but not a good one for discrimination. When the images are taken under large illumination variations, its top eigenvectors mainly model the external illumination variation instead of the identity information. This is a commonly known drawback of PCA and 2DPCA for face recognition with significant illumination changes.

Figure 4.6: Recognition rate of U2DFDA and B2DFDA compared with the other subspace methods on the Yale face database B

Another observation is that kernel Fisher discriminant analysis (KFDA [59] or KDDA [66]) has no consistent excellent performance as 2DFDA. Although KFDA is a little better than 2DFDA on ORL face database when the number of training samples for each subject is larger than 3, the performance of KFDA or KDDA degrades sharply on Yale face database B. The Gaussian RBF kernel is adopted in the K2DPCA and KB2DFDA in all the experiments, the optimal results are achieved when the width, δ , of the kernel is 2.718. From the experiments, we find that KB2DFDA is the best of all the algorithms in recognition performance on ORL database and UMIST database, as shown in Table 4.3 and 4.4.

The fourth experiment is a comparison between the optimal results by 2DFDA and the optimal ones by Fisherface [30]. In Fisherface, the PCA subspace dimension should be $(N - C)$. However, Figure 4.7 shows that the optimal result does not appear at the 120th ($40 \times 4 - 40$) dimension of PCA subspace when there are 4 training samples for each subject. For a fair comparison of 2DFDA and FDA, we design this experiment.

Table 4.3: The comparison of performance on *ORL* database: the first row indicates the number of training sample for each person and the first column indicates the different methods used.

	1	2	3	4	5
KPCA	69.5%	82.5%	88.8%	92.1%	94.2%
LDA	——-	75.8%	87.0%	90.1%	91.7%
KLDA	——-	85.5%	92.2%	95.6%	97.5%
2DPCA	72.5%	84.5%	89.9%	93.1%	95%
K2DPCA	74.5%	86.9%	92.0%	94.6%	96.2%
N-LDA	——-	74.3%	82.9%	87.05%	88.7%
D-LDA	——-	80.6%	85.6%	89.5%	91.7%
KDDA	——-	85.0%	88.6%	92.8%	96.0%
DCV	——-	74.3%	82.9%	87.1%	88.7%
U2DFDA	——-	87.4%	92.5%	95.1%	96.3%
B2DFDA	——-	87.7%	92.7%	95.2%	96.8%
KB2DFDA	——-	89.9%	95.0%	97.2%	98.5%

Figure 4.7: Performance of Fisherface with different dimension of the PCA subspace when the number of training sample is 4 for each subject in the ORL database

Figure 4.8: Optimal performance of FDA compared with optimal performance of U2DFDA and B2DFDA on the ORL database

Figure 4.8 shows the optimal results from 2DFDA and Fisherface on ORL database. Figure 4.9 shows the results on Yale face database B. The results on ORL database shows that the highest recognition rate achieved by B2DFDA surpasses that of FDA by %3 to %10 for small sample case. However, if the number of training samples for each subject is larger than 7, the highest recognition rate achieved by FDA will be higher than that of U2DFDA. The experiment on Yale face database B gives the similar conclusion.

The fifth experiment is the experimental comparison between the performance of U2DFDA and B2DFDA with that of Ye's 2DLDA method [67]. Our experimental results on ORL face database, Figure 4.10, and Yale face database B, Figure 4.11, is in accordance with our previous analysis, i.e., the performance of Ye's 2DLDA is worse than our U2DFDA. Ye's 2DLDA is essentially the re-discriminant analysis (re-projection) of those discriminant features obtained from our U2DFDA, and this re-projection is based on incomplete image information. Therefore, from the perspective of information theory, Ye's iterative optimization scheme is performed at the price of decreasing the available information.

Figure 4.9: Optimal performance of FDA compared with optimal performance of U2DFDA and B2DFDA on the Yale face database B

Figure 4.10: The comparison of performance of B2DFDA and U2DFDA with Ye's 2DLDA on ORL face database

Figure 4.11: The comparison of performance of B2DFDA and U2DFDA with Ye’s 2DLDA on Yale face database B

4.6.2   Performance Comparison on UMIST Database

This section gives a detailed comparison between 2DFDA and the state-of-the-art face recognition algorithms when there exist large pose variations in the face images. This experiment aims to compare the generalization ability of the 2DFDA with that of the other methods. Therefore, we fixed several training sets in the following way. Three experiments with different number of training samples (2, 3, 6 respectively) are conducted to test the generalization ability of our methods. When the number of training samples for each individual is 2, we select {#5, #14} face images of each subject for training, the remaining for test. When the number of training samples is 3 for each subject, six groups are selected for training, i.e., 1{#1, #7, #13}, 2{#2, #8, #14}, 3{#3, #9, #15}, 4{#4, #10, #16}, 5{#5, #11, #17} and 6{#6, #12, #18}, the remaining images corresponding to each group are used to test. When the number of training samples is 6 for each subject, three groups are selected for training, i.e., 1{#1, #4, #7, #10, #13, #16}, 2{#2, #5, #8, #11, #14, #17} and 3{#3, #6, #9, #12, #15, #18}.

From the experimental results on UMIST, Table 4.4, we find that 2DFDA performs

Table 4.4: The comparison of performance on *UMIST* database: the first row indicates the training images used for each person and the first column indicates the different methods used.

	#5, #14	#1, #7, #13	#2, #8, #14	#3, #9, #15	#4, #10, #16	#5, #11, #17	#6, #12, #18
KPCA	80.9%	86.0%	87.0%	91.0%	92.0%	89.3%	87.3%
LDA	77.5%	90.0%	91.3%	95.0%	96.3%	94.3%	91.7%
KLDA	92.5%	94.7%	96.7%	98.3%	99.0%	98.0%	97.3%
2DPCA	90.3%	91.0%	93.0%	95.0%	95.0%	93.7%	92.3%
KDDA	87.8%	94.0%	96.0%	95.7%	97.3%	95.7%	95.7%
DCV	84.1%	89.7%	93.7%	97.7%	94.7%	92.7%	88.0%
K2DPCA	92.7%	94.0%	94.3%	95.7%	97.0%	95.7%	94.0%
U2DFDA	89.3%	93.7%	96.7%	96.7%	96.7%	95.0%	94.0%
B2DFDA	91.3%	96.3%	96.7%	97.3%	97.0%	95.7%	96.0%
KB2DFDA	93.8%	97.5%	97.6%	98.4%	98.6%	96.9%	97.8%

better than the linear subspace method. U2DFDA is comparable to KDDA method and B2DFDA is even better than KDDA. In addition, B2DFDA is comparable to KFDA.

4.7 Concluding Remarks

A 2DFDA framework is proposed to solve the SSS problem in face recognition. It has been analyzed in this chapter that the linear 2DFDA performed on the image matrices is essentially the clustered FLD performed on the rows of all the training samples. Therefore, the enlarged feature set and decreased feature dimension can deal with the Curse of Dimensionality and Small Sample Size problem. The key advantage of the linear 2DFDA over the existing FDA based algorithms is that the within-class scatter matrix is not singular. The extensive experimental results show that the 2DFDA framework outperforms the state-of-the-art linear subspace methods. In addition, the kernel 2DFDA is essentially the clustered KFDA performed on rows/columns if each row/column of the image matrix is viewed as a computational

4.7   Concluding Remarks

97

---

unit. The performance of K2DFDA is better than KFDA in recognition accuracy.

Chapter 5

Discriminant Low-dimensional Subspace Analysis

5.1 Introduction

Linear Discriminant Analysis [56] is a well-known scheme for feature extraction and dimension reduction. It has been used widely in many applications such as face recognition [30, 31], image retrieval [28], text classification [50], micro-array data classification [51], etc. Classical LDA projects the data onto a low-dimensional vector space such that the ratio of the between-class scatter to the within-class scatter is maximized. The optimal transformation can be readily computed by solving a generalized eigenvalue problem.

However, the intrinsic limitation of classical LDA is that its objective function requires the within-class covariance matrix to be nonsingular. For many applications, all scatter matrices in question can be singular since the data vectors lie in a high-dimensional space. In general, the feature dimension far exceeds the number of data samples. To overcome this problem, many approaches have been proposed such as pseudo-inverse LDA [52, 53], PCA+LDA [30], Null-space based LDA (N-LDA) [38], Direct LDA (D-LDA) [39], Random-Subspace LDA [41], Dual-space LDA [43] and

regularized LDA [54, 55].

One common property of all the LDA techniques mentioned above is that the 2D image matrices must be transformed into 1D image vectors before feature extraction. 2D Principal Component Analysis (2DPCA) [9] and 2D Fisher Discriminant Analysis (2DFDA) [139, 140] have been proposed for face recognition. Different from conventional PCA or FLD where data are represented as vectors, 2DPCA and 2DFDA adopt the matrix-based data representation model. That is, the image matrix does not need to be transformed into a vector beforehand. Instead, the covariance matrix is evaluated directly using the 2D image matrices. Therefore, the size of the covariance matrix in question is much smaller and the performance of 2DPCA is better than PCA on face recognition, image compression and retrieval. In contrast to the $\mathbf{S}_b$ and $\mathbf{S}_w$ of conventional FLD, the covariance matrices obtained by 2DFDA are generally not singular.

To overcome the Small Sample Size problem in pursuit of the optimal projection directions in the whole face-image space, a framework of low-dimensional Fisher discriminant subspace analysis is proposed in this chapter for face recognition with each subject having small number of training images. In this framework, it is mathematically proven that the null space of the total covariance matrix, $\mathbf{S}_t$, is useless for recognition. Therefore, a low-dimensional Fisher discriminant analysis is developed in two different ways by discarding the null space of $\mathbf{S}_t$. Two different algorithms are thus proposed. A Unified Linear Discriminant Analysis (ULDA) is proposed in the first algorithm to extract the discriminant information from three subspaces of the range space of $\mathbf{S}_t$. The second algorithm, called Modified Linear Discriminant Analysis (MLDA), can avoid the numerical problem in conventional LDA by adopting a *Deduction*-form Fisher criterion in place of the *Quotient*-form one. Furthermore, the Kernel Modified Fisher Discriminant Analysis (KMFDA) is also developed to extract the nonlinear discriminant vectors, and the numerical problem existing in conventional KFDA [59, 65] does not exist any more in KMFDA.

The rest of the chapter is organized as follows: Section 5.2 proposes the ULDA for

the first algorithm. The MLDA and KMFDA with new Fisher criterion are proposed in section 5.3 and 5.4 respectively. Experiments and discussions are presented in Section 5.5. We draw the conclusion in Section 5.6.

5.2 Unified Discriminant Subspace Analysis

The conventional two-stage LDA-based algorithms, such as Fisherface [30], Null-space based LDA [38, 40], Direct-LDA [39], etc., all lose some useful information to ensure the nonsingularity of the within-class covariance matrix by discarding part of the discriminant subspaces. In this section, our task is to unify the specific discriminant analysis methods by splitting subspace and fusing extracted features from them.

It is well known that the data are projected onto a lower-dimensional vector space through the LDA transformation such that the ratio of the between-class scatter to the within-class scatter is maximized, thus achieving maximum discrimination. The optimal projection (transformation), $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{L-1}]$, can be readily computed by solving a generalized eigenvalue problem, where $\mathbf{w}_i, i = 1, \dots, L - 1$ satisfy the following conventional Fisher criterion,

$$\mathbf{w} = \arg \max \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}} \quad (5.1)$$

If $\mathbf{S}_w$ is invertible, $\mathbf{w}^T \mathbf{S}_w \mathbf{w}$ is always positive for every nonzero $\mathbf{w}$ since $\mathbf{S}_w$ is positive definite. In such a case, Equation (5.1) can be used directly to extract a set of optimal discriminant projection vectors. However, it is almost impossible in real-world application, such as face recognition, that $\mathbf{S}_w$ is of full rank. Therefore, there always exist vectors (these vectors are from $\text{Null}(\mathbf{S}_w)$) making $\mathbf{w}^T \mathbf{S}_w \mathbf{w}$ be zero. These vectors turn out to be very effective for classification if they satisfy $\mathbf{w}^T \mathbf{S}_b \mathbf{w} > 0$ meantime. This is the idea of Null-space based LDA methods [38, 40]

where they adopt the modified criterion as follows,

$$\mathbf{w} = \arg \max_{\mathbf{w}^T \mathbf{S}_w \mathbf{w} = 0} \mathbf{w}^T \mathbf{S}_b \mathbf{w} \quad (5.2)$$

Before the detailed description of Algorithm 1, we provide the theoretical foundation to it.

Lemma 5.1: The null space of $\mathbf{S}_t$ is the common null space of both $\mathbf{S}_b$ and $\mathbf{S}_w$.

Proof: Let $\mathbf{W}_{null}$ be the null space of $\mathbf{S}_t$, that is,

$$\mathbf{W}_{null}^T \mathbf{S}_t \mathbf{W}_{null} = 0 \quad (5.3)$$

It is trivial to check that $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$. Therefore, we have,

$$\mathbf{W}_{null}^T (\mathbf{S}_b + \mathbf{S}_w) \mathbf{W}_{null} = 0 \quad (5.4)$$

Since both $\mathbf{S}_b$ and $\mathbf{S}_w$ are positive semi-definite, we have,

$$\mathbf{W}_{null}^T \mathbf{S}_b \mathbf{W}_{null} = 0 \quad \text{and} \quad \mathbf{W}_{null}^T \mathbf{S}_w \mathbf{W}_{null} = 0 \quad (5.5)$$

Therefore, we can conclude that the null space of $\mathbf{S}_t$ is the common null space of $\mathbf{S}_b$ and $\mathbf{S}_w$. $\square$

Let $\mathcal{F}$ be the vector space where all the face-image vectors lie. Since $\mathbf{S}_t$ is symmetric and positive semi-definite, its eigenvectors that correspond to the nonzero eigenvalues forms a set of orthonormal basis for $\mathcal{F}$. Generally, $\text{Rank}(\mathbf{S}_t) = n - 1$. Suppose $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{n-1}$ are the set of eigenvectors, then $\text{Range}(\mathbf{S}_t) = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{n-1}\}$ and $\mathcal{F} = \text{Range}(\mathbf{S}_t) \oplus \text{Null}(\mathbf{S}_t)$. For any vector $\mathbf{z} \in \mathcal{F}$, $\mathbf{z}$ can be uniquely represented by $\mathbf{z} = \mathbf{g} + \mathbf{h}$ with $\mathbf{g} \in \text{Range}(\mathbf{S}_t)$ and $\mathbf{h} \in \text{Null}(\mathbf{S}_t)$.

Lemma 5.2: If matrix $\mathbf{A}$ is positive definite, $\mathbf{x}^T \mathbf{A} \mathbf{x} = 0$ if and only if $\mathbf{A} \mathbf{x} = 0$.

Proof: It is trivial to see that if $\mathbf{Ax} = 0$, $\mathbf{x}^T \mathbf{Ax} = 0$. Therefore, what we need is only to prove that $\mathbf{x}^T \mathbf{Ax} = 0$ will result in $\mathbf{Ax} = 0$. Since $\mathbf{A}$ is positive definite, it must have a positive square root $\mathbf{R}$ such that $\mathbf{A} = \mathbf{R}^2$. Thus, $\langle \mathbf{Rx}, \mathbf{Rx} \rangle = \langle \mathbf{Ax}, \mathbf{x} \rangle = \mathbf{x}^T \mathbf{Ax} = 0$. Therefore, $\mathbf{Rx} = 0$. We can further derive that $\mathbf{Ax} = \mathbf{R}(\mathbf{Rx}) = 0$. $\square$

Theorem 5.1: The discriminative features can only be extracted from the range space of $\mathbf{S}_t$.

Proof: For any vector $\mathbf{h} \in \text{Null}(\mathbf{S}_t)$, we have $\mathbf{h}^T \mathbf{S}_t \mathbf{h} = 0$. From Theorem 5.1, we have $\mathbf{h}^T \mathbf{S}_b \mathbf{h} = 0$. Since $\mathbf{S}_b$ is positive, according to Lemma 5.1, we have $\mathbf{S}_b \mathbf{h} = 0$. Hence, for any vector $\mathbf{z} \in \mathcal{F} = \mathbf{g} + \mathbf{h}$, where $\mathbf{g} \in \text{Range}(\mathbf{S}_t)$ and $\mathbf{h} \in \text{Null}(\mathbf{S}_t)$,

$$\begin{aligned} \mathbf{z}^T \mathbf{S}_b \mathbf{z} &= (\mathbf{g} + \mathbf{h})^T \mathbf{S}_b (\mathbf{g} + \mathbf{h}) \\ &= \mathbf{g}^T \mathbf{S}_b \mathbf{g} + 2\mathbf{g}^T \mathbf{S}_b \mathbf{h} + \mathbf{h}^T \mathbf{S}_b \mathbf{h} \\ &= \mathbf{g}^T \mathbf{S}_b \mathbf{g} \end{aligned}$$

Similarly, we also have

$$\begin{aligned} \mathbf{z}^T \mathbf{S}_w \mathbf{z} &= (\mathbf{g} + \mathbf{h})^T \mathbf{S}_w (\mathbf{g} + \mathbf{h}) \\ &= \mathbf{g}^T \mathbf{S}_w \mathbf{g} + 2\mathbf{g}^T \mathbf{S}_w \mathbf{h} + \mathbf{h}^T \mathbf{S}_w \mathbf{h} \\ &= \mathbf{g}^T \mathbf{S}_w \mathbf{g} \end{aligned}$$

Therefore, the conventional Fisher criterion in Equation (5.1) and modified criterion of Null-space based LDA in Equation (5.2) will be converted into

$$\begin{aligned} \mathbf{g} &= \arg \max_{\mathbf{g} \in \text{Range}(\mathbf{S}_t)} \frac{(\mathbf{g} + \mathbf{h})^T \mathbf{S}_b (\mathbf{g} + \mathbf{h})}{(\mathbf{g} + \mathbf{h})^T \mathbf{S}_w (\mathbf{g} + \mathbf{h})} \\ &= \arg \max_{\mathbf{g} \in \text{Range}(\mathbf{S}_t)} \frac{\mathbf{g}^T \mathbf{S}_b \mathbf{g}}{\mathbf{g}^T \mathbf{S}_w \mathbf{g}} \end{aligned}$$

and

$$\begin{aligned}
\mathbf{g} &= \arg \max_{(\mathbf{g}+\mathbf{h})^T \mathbf{S}_w (\mathbf{g}+\mathbf{h})=0} (\mathbf{g}+\mathbf{h})^T \mathbf{S}_b (\mathbf{g}+\mathbf{h}) \\
&= \arg \max_{\mathbf{g}^T \mathbf{S}_w \mathbf{g}=0} \mathbf{g}^T \mathbf{S}_b \mathbf{g}
\end{aligned}$$

Therefore, we can draw the conclusion that the discriminant projection vectors can be extracted from $\text{Range}(\mathbf{S}_t) = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{n-1}\}$, i.e., $\text{Null}(\mathbf{S}_t)$ has little contribution to recognition and can be safely discarded without losing discriminant information. $\square$

Based on Theorem 5.1, the face-image vectors are all projected onto a low-dimensional (n-1) space, called $\mathcal{Q}$, determined by the eigenvectors, $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{n-1}]$, of $\mathbf{S}_t$ corresponding to those nonzero eigenvalues. Therefore, the between-class and within-class covariance matrices are transformed into $\hat{\mathbf{S}}_b = \mathbf{U}^T \mathbf{S}_b \mathbf{U}$ and $\hat{\mathbf{S}}_w = \mathbf{U}^T \mathbf{S}_w \mathbf{U}$ respectively in $\mathcal{Q}$. The transformed total covariance matrix $\hat{\mathbf{S}}_t = \hat{\mathbf{S}}_b + \hat{\mathbf{S}}_w$. It is easy to check that $\text{Rank}(\hat{\mathbf{S}}_t) = n - 1$, $\text{Rank}(\hat{\mathbf{S}}_b) = L - 1$ and $\text{Rank}(\hat{\mathbf{S}}_w) = n - L$. The standard LDA method remains inapplicable since $\hat{\mathbf{S}}_w$ is still singular in $\mathcal{Q}$. To extract the discriminant information from $\mathcal{Q}$ as complete as possible, we use a unified discriminant subspace analysis method in a divide and conquer way. Our strategy is to split the $\mathcal{Q}$ space first, then a specific algorithm will be adopted in the individually split subspace of $\mathcal{Q}$ to deal with different situations. We split the $\mathcal{Q}$ space in two ways. The first way is to split $\mathcal{Q}$ into two subspaces: $\text{Null}(\hat{\mathbf{S}}_w)$ and $\text{Range}(\hat{\mathbf{S}}_w)$. The other way is to split $\mathcal{Q}$ into another two subspaces: $\text{Null}(\hat{\mathbf{S}}_b)$ and $\text{Range}(\hat{\mathbf{S}}_b)$. Therefore, we have four subspaces available for feature extraction. However, $\text{Null}(\hat{\mathbf{S}}_b)$ is not used considering its little contribution to the discriminant feature extraction. Then we use three different LDA methods for the three subspaces, i.e., the standard LDA is used in the $\text{Range}(\hat{\mathbf{S}}_w)$, Null-space based LDA is used in the $\text{Null}(\hat{\mathbf{S}}_w)$ and Direct-LDA is used in the $\text{Range}(\hat{\mathbf{S}}_b)$. Generally, we have the following steps for the ULDA algorithm. They are listed in Table 5.1.

After obtaining the $\mathbf{W}_{opt}^{LDA}$, $\mathbf{W}_{opt}^{NLDA}$ and $\mathbf{W}_{opt}^{DLDA}$, we can extract three $(L - 1)$ -

Table 5.1: The algorithm of Unified Linear Discriminant Analysis**Algorithm 1:** The ULDA Algorithm

1. Compute the mean of each class, $\mathbf{m}_i$, and the mean of all the classes, $\mathbf{m}$. Construct $\mathbf{BM} = [(\mathbf{m}_1 - \mathbf{m}), (\mathbf{m}_2 - \mathbf{m}), \dots, (\mathbf{m}_L - \mathbf{m})]$, let $\tilde{\mathbf{S}}_b = \mathbf{BM}^T \mathbf{B} \mathbf{M}$
2. Construct $\mathbf{WM} = [(\mathbf{x}_1^1 - \mathbf{m}_1), (\mathbf{x}_1^2 - \mathbf{m}_1), \dots, (\mathbf{x}_L^{n_L} - \mathbf{m}_L)]$. Let $\tilde{\mathbf{S}}_w = (\mathbf{WM})^T \mathbf{W} \mathbf{M}$. Similarly, construct $\mathbf{TM} = [(\mathbf{x}_1^1 - \mathbf{m}), (\mathbf{x}_1^2 - \mathbf{m}), \dots, (\mathbf{x}_L^{n_L} - \mathbf{m})]$. Let $\tilde{\mathbf{S}}_t = (\mathbf{TM})^T \mathbf{T} \mathbf{M}$
3. Compute the eigenvectors of $\mathbf{S}_t$ corresponding to the top $(n-1)$ eigenvalues as $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{n-1}]$ using the trick proposed in Eigenface method
4. Compute the $\hat{\mathbf{S}}_b = \mathbf{U}^T \mathbf{B} \mathbf{M} (\mathbf{B} \mathbf{M})^T \mathbf{U}$ and $\hat{\mathbf{S}}_w = \mathbf{U}^T \mathbf{W} \mathbf{M} (\mathbf{W} \mathbf{M})^T \mathbf{U}$
5. Perform standard LDA in $\text{Range}(\hat{\mathbf{S}}_w)$: Get $\mathbf{W}^{LDA} = [\mathbf{w}_1^{LDA}, \dots, \mathbf{w}_{L-1}^{LDA}]$
6. Perform Null-space based LDA in $\text{Null}(\hat{\mathbf{S}}_w)$: Get $\mathbf{W}^{NLDA} = [\mathbf{w}_1^{NLDA}, \mathbf{w}_2^{NLDA}, \dots, \mathbf{w}_{L-1}^{NLDA}]$
7. Perform Direct-LDA in $\text{Range}(\hat{\mathbf{S}}_b)$: Get $\mathbf{W}^{DLDA} = [\mathbf{w}_1^{DLDA}, \mathbf{w}_2^{DLDA}, \dots, \mathbf{w}_{L-1}^{DLDA}]$.
8. Construct $\mathbf{W}_{opt}^{LDA}$, $\mathbf{W}_{opt}^{NLDA}$ and $\mathbf{W}_{opt}^{DLDA}$: $\mathbf{W}_{opt}^{LDA} = \mathbf{U} * \mathbf{V}_{Range} * \mathbf{W}^{LDA}$, $\mathbf{W}_{opt}^{NLDA} = \mathbf{U} * \mathbf{V}_{Null} * \mathbf{W}^{NLDA}$, $\mathbf{W}_{opt}^{DLDA} = \mathbf{U} * \mathbf{P}_{Range} * \mathbf{W}^{DLDA}$

dimensional discriminant feature vectors for any given sample $\mathbf{x}$ via three linear transformation, $\mathbf{y}^{LDA} = (\mathbf{W}_{opt}^{LDA})^T \mathbf{x}$, $\mathbf{y}^{NLDA} = (\mathbf{W}_{opt}^{NLDA})^T \mathbf{x}$ and $\mathbf{y}^{DLDA} = (\mathbf{W}_{opt}^{DLDA})^T \mathbf{x}$. Therefore, it is possible to fuse them in a decision level. Here, we propose a fusion strategy and use it for classification in this paper.

Suppose the distance between any two data, $\mathbf{x}_i$ and $\mathbf{x}_j$, is given by the Euclidean distance, i.e., $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. Let us denote a pattern $\mathbf{y} = [\mathbf{y}^{LDA}, \mathbf{y}^{NLDA}, \mathbf{y}^{DLDA}]$. The summed normalized-distance between $\mathbf{y}$ and those of the training samples $\mathbf{y}_i = [\mathbf{y}_i^{LDA}, \mathbf{y}_i^{NLDA}, \mathbf{y}_i^{DLDA}]$, $i = 1, 2, \dots, n$, is given by,

$$d(\mathbf{y}, \mathbf{y}_i) = d^{LDA} + \alpha * d^{NLDA} + \beta * d^{DLDA} \quad (5.6)$$

where α and β are used to adjust the relative importance of each distance, $d^{LDA} = \frac{\|\mathbf{y}^{LDA} - \mathbf{y}_i^{LDA}\|}{\sum_{j=1}^n \|\mathbf{y}^{LDA} - \mathbf{y}_j^{LDA}\|}$, $d^{NLDA} = \frac{\|\mathbf{y}^{NLDA} - \mathbf{y}_i^{NLDA}\|}{\sum_{j=1}^n \|\mathbf{y}^{NLDA} - \mathbf{y}_j^{NLDA}\|}$ and $d^{DLDA} = \frac{\|\mathbf{y}^{DLDA} - \mathbf{y}_i^{DLDA}\|}{\sum_{j=1}^n \|\mathbf{y}^{DLDA} - \mathbf{y}_j^{DLDA}\|}$.

In classification, the nearest-neighborhood classifier is used.

5.3 Modified Linear Discriminant Analysis

In this section, a Modified Linear Discriminant Analysis (MLDA) algorithm is proposed. In the conventional LDA, the Fisher optimization criterion is in a *Quotient* form. From the analysis mentioned above, the *Quotient* form will induce the numerical problem when there only exist small number of gallery images for each subject in the databases. To overcome this problem and meanwhile to be consistent with the Fisher rule (maximizing the between-class scatter and minimizing the within-class scatter), a modified Fisher optimization criterion in the *deduction* form is proposed. That is,

$$\mathbf{W}_{fld} = \arg \max_{\mathbf{W}} \text{tr}\{\mathbf{W}^T(\mathbf{S}_b - \mu\mathbf{S}_w)\mathbf{W}\} \quad (5.7)$$

where $\text{tr}(\cdot)$ is the trace operation. It is trivial to check that this modified criterion is consistent with the Fisher rule.

To get the $\mathbf{W}_{fld}$, we only need to get the solutions to the following eigenvalue problem,

$$(\mathbf{S}_b - \mu\mathbf{S}_w)\mathbf{w}_i = \lambda_i\mathbf{w}_i \quad (5.8)$$

However, it is difficult to directly solve Equation (5.8) where $\mathbf{S}_b$ and $\mathbf{S}_w$ are pretty large matrix. Fortunately, we have analyzed in Theorem 5.2 that the null space of $\mathbf{S}_t$ is useless for recognition, therefore, the $\text{Null}(\mathbf{S}_t)$ is discarded beforehand. The proposed MLDA is listed as Algorithm 2 in Table 5.2.

Table 5.2: The algorithm of Modified Linear Discriminant Analysis

---

Algorithm 2: The MLDA Algorithm

Input: All gallery data $\mathbf{A}$

Output: Discriminant vectors $\{\mathbf{w}_i\}$'s of MLDA

1. Compute the eigenvectors of $\mathbf{S}_t$ corresponding to the top $(n-1)$ eigenvalues as $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{n-1}]$ using the method in Algorithm 1
  2. Compute the $\hat{\mathbf{S}}_b = \mathbf{U}^T \mathbf{S}_b \mathbf{U}$ and $\hat{\mathbf{S}}_w = \mathbf{U}^T \mathbf{S}_w \mathbf{U}$
  3. Solve the eigen-equation: $(\hat{\mathbf{S}}_b - \mu\hat{\mathbf{S}}_w)\hat{\mathbf{w}}_i = \lambda_i\hat{\mathbf{w}}_i$. Let $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \dots, \hat{\mathbf{w}}_l]$, where l generally equals $L - 1$
  4. $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{L-1}] \leftarrow \mathbf{U}\hat{\mathbf{W}}$
-

Once given the projection matrix $\mathbf{W}_{fld}$, the projection of any data, $\mathbf{x}$, onto this subspace is given by $\mathbf{y} = \mathbf{W}_{fld}^T \mathbf{x}$, the nearest-neighborhood classifier is used for classification.

5.4 Kernel Modified Fisher Discriminant Analysis

By mapping the original data into the reproducing kernel Hilbert space (RKHS) using a kernel trick, the kernelized subspace analysis has demonstrated the superiority over their linear counterpart [59, 65]. In this section, a Kernel Modified Fisher Discriminant Analysis (KMFDA) is proposed to extend the MFLD to the higher-dimensional feature space. In the kernel space, high-order statistics among the input data is investigated and the discriminant features nonlinearly related to the input data is extracted. Compared with the GDA [65] and Kernel Fisherface [59], it has two advantages. Firstly, no numerical problem exists in KMFDA any more. Secondly, it does not discard any valuable information for recognition. These advantages will be demonstrated in the later part of this section.

The idea of KMFDA is to solve the problem of MFLD in an implicit higher-dimensional (maybe infinite-dimensional) feature space $\mathcal{H}$ constructed by a non-linear mapping, $\Phi : \mathbf{x} \in \mathcal{R}^N \mapsto \Phi(\mathbf{x}) \in \mathcal{H}$. In implementation, the mapping, Φ , is implicit and does not need to be computed explicitly, instead, it is embedded by computing the inner product of two vectors in $\mathcal{H}$ with a kernel function, $k(x, y) = \langle \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \rangle$.

Let $\mathbf{m}_i^\Phi$ be the mean of the gallery image vectors of the i -th class and $\mathbf{m}^\Phi$ be the mean of the total gallery image vectors in $\mathcal{H}$, i.e., $\mathbf{m}_i^\Phi = \frac{1}{n_i} \sum_{k=1}^{n_i} \{\Phi(\mathbf{x}_k) | \Phi(\mathbf{x}_k) \in \mathbf{X}_i\}$ and $\mathbf{m}^\Phi = \frac{1}{n} \sum_{i=1}^L \sum_{j=1}^{n_i} \{\Phi(\mathbf{x}_j) | \Phi(\mathbf{x}_j) \in \mathbf{X}_i\}$. Let $\mathbf{S}_b^\Phi$ and $\mathbf{S}_w^\Phi$ be the between-class scatter matrix and within-class scatter matrix respectively in $\mathcal{H}$. Then $\mathbf{S}_b^\Phi$ and $\mathbf{S}_w^\Phi$

are represented as:

$$\mathbf{S}_w^\Phi = \sum_{i=1}^L \sum_{\Phi(\mathbf{x}_k) \in \mathcal{X}_i} (\Phi(\mathbf{x}_k) - \mathbf{m}_i^\Phi)(\Phi(\mathbf{x}_k) - \mathbf{m}_i^\Phi)^T \quad (5.9)$$

$$\mathbf{S}_b^\Phi = \sum_{i=1}^L n_i (\mathbf{m}_i^\Phi - \mathbf{m}^\Phi)(\mathbf{m}_i^\Phi - \mathbf{m}^\Phi)^T \quad (5.10)$$

Then to perform the MLDA in $\mathcal{H}$ is equivalent to maximizing the following Fisher discriminant function.

$$\mathbf{W}_{kfd}^\Phi = \arg \max_{\mathbf{W}^\Phi} \text{tr}\{(\mathbf{W}^\Phi)^T (\mathbf{S}_b^\Phi - \mu \mathbf{S}_w^\Phi) \mathbf{W}^\Phi\} \quad (5.11)$$

where $\mathbf{W}_{kfd}^\Phi = [\mathbf{w}_1^\Phi, \mathbf{w}_2^\Phi, \dots, \mathbf{w}_l^\Phi]$ are the kernel discriminant projection vectors in $\mathcal{H}$.

Because of the high-dimensional feature space, it is impossible to solve Equation (5.11) using the common method as in MLDA. Fortunately, with the aid of kernel trick, i.e., $k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) \rangle$, the nonlinear discriminant component can be extracted.

As any solution $\mathbf{w}_k^\Phi \in \mathcal{H}$ must lie in the span of all the gallery samples in $\mathcal{H}$, i.e., there exist coefficients $\tilde{\alpha}_i, i = 1, 2, \dots, l$, and $\tilde{\alpha}_k = \alpha_k^j, j = 1, 2, \dots, n$, such that

$$\mathbf{w}_k^\Phi = \sum_{i=1}^n \alpha_k^i \Phi(\mathbf{x}_i), \quad k = 1, 2, \dots, l \quad (5.12)$$

Therefore, the projection of the mean of the gallery samples of each class in $\mathcal{H}$, $\mathbf{m}_i^\Phi$, onto $\mathbf{w}_k^\Phi$ can be written as,

$$(\mathbf{w}_k^\Phi)^T \mathbf{m}_i^\Phi = \tilde{\alpha}_k^T \begin{bmatrix} \frac{1}{n_i} \sum_{j=1}^{n_i} k(\mathbf{x}_1, \mathbf{x}_j) \\ \frac{1}{n_i} \sum_{j=1}^{n_i} k(\mathbf{x}_2, \mathbf{x}_j) \\ \vdots \\ \frac{1}{n_i} \sum_{j=1}^{n_i} k(\mathbf{x}_n, \mathbf{x}_j) \end{bmatrix} = \tilde{\alpha}_k^T \mathbf{z}_i \quad (5.13)$$

Similarly, the projection of the mean of the gallery samples of all the classes in $\mathcal{H}$, $\mathbf{m}^\Phi$, onto $\mathbf{w}_k^\Phi$ can be written as,

$$(\mathbf{w}_k^\Phi)^T \mathbf{m}^\Phi = \frac{\tilde{\alpha}_k^T}{n} \begin{bmatrix} \sum_{i=1}^L \sum_{j=1}^{n_i} k(\mathbf{x}_1, \mathbf{x}_j | \mathbf{x}_j \in \mathbf{X}_i) \\ \sum_{i=1}^L \sum_{j=1}^{n_i} k(\mathbf{x}_2, \mathbf{x}_j | \mathbf{x}_j \in \mathbf{X}_i) \\ \vdots \\ \sum_{i=1}^L \sum_{j=1}^{n_i} k(\mathbf{x}_n, \mathbf{x}_j | \mathbf{x}_j \in \mathbf{X}_i) \end{bmatrix} = \tilde{\alpha}_k^T \mathbf{z} \quad (5.14)$$

From Equation (5.13) and Equation (5.14), we can easily obtain the following equations,

$$(\mathbf{w}_k^\Phi)^T \mathbf{S}_b^\Phi \mathbf{w}_k^\Phi = \tilde{\alpha}_k^T \mathbf{K}_b \tilde{\alpha}_k \quad (5.15)$$

where

$$\mathbf{K}_b = \sum_{i=1}^L n_i (\mathbf{z}_i - \mathbf{z})(\mathbf{z}_i - \mathbf{z})^T \quad (5.16)$$

Similarly, we also can get

$$(\mathbf{w}_k^\Phi)^T \mathbf{S}_w^\Phi \mathbf{w}_k^\Phi = \tilde{\alpha}_k^T \mathbf{K}_w \tilde{\alpha}_k \quad (5.17)$$

where

$$\mathbf{K}_w = \sum_{i=1}^L \sum_j^{n_i} (\mathbf{u}_j - \mathbf{z}_i)(\mathbf{u}_j - \mathbf{z}_i)^T \quad (5.18)$$

and

$$\mathbf{u}_j = (k(\mathbf{x}_1, \mathbf{x}_j), k(\mathbf{x}_2, \mathbf{x}_j), \dots, k(\mathbf{x}_n, \mathbf{x}_j))^T \quad (5.19)$$

Thus, the maximization of Equation (5.11) is transformed into the maximization of the following function,

$$\mathbf{J}(\tilde{\alpha}) = \text{tr}\{\tilde{\alpha}^T (\mathbf{K}_b^\Phi - \mu \mathbf{K}_w^\Phi) \tilde{\alpha}\} \quad (5.20)$$

Maximizing Equation (5.20) is equivalent to solving the eigenvalue problem of

$$(\mathbf{K}_b^\Phi - \mu \mathbf{K}_w^\Phi) \tilde{\alpha}_i = \lambda_i \tilde{\alpha}_i \quad (5.21)$$

Therefore, the eigenvectors corresponding to the leading eigenvalues of $\mathbf{K}_b^\Phi - \mu \mathbf{K}_w^\Phi$ are the coefficient-vectors, $\vec{\alpha}_i, i = 1, 2, \dots, l$. The projection of $\Phi(\mathbf{x})$ onto $\mathbf{w}_k^\Phi$ can be represented as,

$$(\mathbf{w}_k^\Phi)^T \Phi(\mathbf{x}) = \sum_{i=1}^n \alpha_k^i k(\mathbf{x}_i, \mathbf{x}) \quad (5.22)$$

5.5 Experimental Results and Discussion

To test the proposed algorithms, ULDA, MLDA and KMFDA, in the situation where there are a large number of subjects in the database and very few training samples are available for each subject, meanwhile, there are pose, illumination and expression variations for the subjects, we construct a large database by combining several existing public databases, i.e., ORL, Yale, YaleB, CMU PIE, UMIST [159], CMU AMP Expression [160], and XM2VTS databases.

The ORL face database contains images from 40 individuals, each providing 10 different images. There exist variations in pose, facial expression and face scale. In our experiment, altogether 100 images for 10 subjects of YaleB face dataset are used (10 mild illumination conditions under the same frontal pose). CMU PIE contains images with pose, illumination and expression variations with each person imaged under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In our experiment, altogether 450 images for 45 subjects are used (10 mild illumination conditions under the same frontal pose). All the images from YaleB and PIE databases are preprocessed and normalized by translation, rotation, and scaling such that the two outer eye corners are in fixed positions.

For UMIST database, 200 images of 20 persons with each subject having 10 samples are used to ensure that face appearance changes from profile to frontal orientation with a step of about 10° separation.

In the CMU face expression database, the images are collected only with expression variations. There are 13 subjects with each one has 75 images showing expression

variations. In our experiments, 130 images are used for 13 subjects with each subject has 10 different expressions.

We use the 3D VRML face model to render 10 images for 30 subjects in XM2VTS database, therefore, 300 images are used for XM2VTS database. For Yale database, we use 200 images for 20 subjects with 10 images for each subject. Some sample images are shown in Fig.5.1.

Figure 5.1: Sample images in the combined database

Therefore, we have 1730 images for altogether 173 subjects with each person having 10 images in the large combined database. All the images are normalized into the same size of 92×112 . It has been noticed that the face images from Yale face database are not normalized in the same way as the face images from the other face databases, and this will affect the recognition performance. However, we believe that the effect is in the same way for all the compared algorithms because all the compared algorithms are trained on the same training set and tested on the same test set. To verify the claim, a simple test is made. The test is to use the completely normalized combined face database. We then compare the recognition performance with the current recognition performance. We test four algorithms

Figure 5.2: Experimental results on the combined database

(MLDA, 2DFDA, Fisherface, and N-LDA). The results are that the recognition performance of all the four algorithms degrade (accuracy decreasing 1.8%, 1.5%, 2.6% and 2.4% correspondingly). This test is consistent with our analysis that the un-normalized face images will affect the recognition performance of the compared subspace methods in the same direction (accuracy all decreasing or all increasing).

However, for the convenience of comparison of all the algorithms, the comparison of the performance of the three proposed algorithms with the state-of-the-art methods, e.g., N-LDA, D-LDA, 2DFDA, KFDA etc, is still made on the original combined face database. The comparison results are shown in Fig.5.2.

5.5.1 Random Grouping of Training and Testing Sets

To test the recognition performance with different training numbers, k ($2 \leq k \leq 5$) images of each subject are randomly selected for training and the remaining $10-k$

images of each subject for testing. For each number k , 50 times of random selections are performed the combined database. The final recognition rate is the average of all.

5.5.2 Comparison of Recognition Performance

In Fisherface, N-LDA and D-LDA, the dimension of the feature vector for classification is all reserved to $L - 1$. In Kernel Fisherface, the Gaussian RBF kernel is used. The reserved dimension for classification is $L - 1$. The optimal kernel width for classification is about $4 \sim 5$. For 2DPCA and 2DFDA, the results shown in Fig.5.2 are all from their optimal one, the reserved feature matrix is about 112×3 or 112×4 . For the proposed algorithms in this paper, the dimensions reserved for classification are all $L - 1$.

From the experimental comparison, we may find that 2DPCA can achieve a much better performance than Fisherface, N-LDA and D-LDA. The underlying reason is that the images in the combined database contain only mild illumination and pose variation, 2DPCA can achieve a good generalization. In the meantime, the number of training samples is small, the performance of the LDA-based algorithm (Fisherface, N-LDA and D-LDA) often degrades sharply (making the within-class covariance matrix nonsingular at the price of losing part of useful subspaces). Therefore, their performance is often inferior to that of the 2DPCA. In addition, consistent with the experimental results of the previous papers, we do find that N-LDA and D-LDA is better than Fisherface method in recognition accuracy. N-LDA is superior to D-LDA. 2DFDA can have better performance than 2DPCA since the SSS problem does not exist any more in both of them, and 2DPCA is good at representation rather than discrimination. The proposed MLDA can bypass the SSS problem by avoiding the direct use of within-class covariance matrix, and it is also consistent with the major idea of Fisher criterion. For this reason, it has better discriminant ability, especially when there are only few training samples for each subject, over

2DPCA, Fisherface, N-LDA and D-LDA.

The proposed ULDA takes full advantage of all the discriminant information from the possible useful subspace instead of using only one of them. Therefore it can achieve higher recognition rate than the other LDA-based algorithms. In addition, MLDA and ULDA can both achieve better recognition performance than 2DFDA, ULDA is better than MLDA and KFDA. KMFDA can avoid the numerical problem in standard KFDA, it achieves highest recognition rate among all the listed algorithms.

5.5.3 The Effect of α and β on Recognition rate

To show the effect of different α and β on recognition accuracy, we use one specific example to illustrate it where only two images are available (the 2-th and 6-th images of each subject) for training. We change the values of α and β both from 0 ~ 4.2 with a step of 0.2, Fig.5.3 shows the experimental results, where (a) is viewed from above and (b) is the side view. From Fig.5.3, we find that the recognition performance is the best when α and β are at 1.8.

5.5.4 The Effect of μ on Recognition Rate

In MLDA, the performance will be different if μ takes different values. To show the effect of different μ on recognition accuracy, a coarse adjustment of μ is followed by a fine adjustment. The number of training samples for each subject is 2, 3 and 4 for this experiment. For the coarse adjustment, the range is from 0 to 500 with a step of 5. For the fine adjustment, the range is from 0 to 20 with a step of 0.2, Fig.5.4 shows the experimental results, where the results in the first row is for the coarse adjustment and the second row denotes the fine adjustment. From Fig.5.4, we find it can achieve the highest recognition rate when the μ is 5, 2.5 and 2.5 respectively in this experiment.

5.6 Concluding Remarks

A framework of low-dimensional Fisher discriminant analysis is developed in two different ways by discarding the null space of $\mathbf{S}_t$. Two different algorithms are thus proposed, i.e., Unified Linear Discriminant Analysis (ULDA) and Modified Linear Discriminant Analysis (MLDA). Furthermore, the Kernel version of MLDA (KMFDA) is also developed to extract the nonlinear discriminant vectors. Experimental results on a large combined face database have demonstrated that the proposed two linear schemes in this framework can both achieve better performance than the other state-of-the-art LDA-based algorithms, and the KMFDA is superior to KFDA in recognition accuracy.

(a)

(b)

Figure 5.3: The effect of α and β on recognition rate

Figure 5.4: The performance of MLDA in recognition with different μ value. (a)-(c): Coarse adjustment of μ with two, three and four training samples respectively for each subject; (d)-(f): Fine adjustment of μ with two, three and four training samples respectively for each subject.

Chapter 6

Ensemble Learning for Diversified Fisher Linear Discriminant

6.1 Introduction

The low-dimensional discriminant subspace analysis has been a hotspot in computer vision area in recent years. Fisherface [30] extracts features from the range space of the within-class scatter matrix. Null-space LDA (N-LDA) methods extract features from the null space of the within-class scatter matrix [38, 40, 46]. Discriminant Common Vectors (DCV) [49] are used for representing each person in the training set of the face database. However, this scheme is essentially a null-space-like method. As stated in [49], the success of this method depends on the size of the null space of the within-class scatter matrix. When the size of the null space is small, recognition rates are expected to be poor, since there will not be sufficient space for obtaining the optimal projection vectors. 2D Fisher Discriminant Analysis (2DFDA) was proposed for face recognition [140] with matrix-based data representation model.

Contemporary to the single-classifier approaches, multiple classifier systems have also been widely investigated in pattern recognition area over the past decade [144–149]. These multiple classifier systems can be generalized into two classes. One

is based on the combination of multiple cues for classification [144, 148] e.g., the face/iris/ear images together with speech are used for person identification. The other is based on the ensemble of multiple weak classifiers with a single cue utilized [146, 147, 149]. The ensemble methods such as random subspace [42], adaboost [142] and bagging [141] have been widely discussed in machine learning area. However, their performance in face recognition has not been studied broadly until recently [41, 44, 45, 47, 48]. In [47], a bagging-like scheme is utilized to generate several subsets of samples from the training data set. LDA is applied to each of the subsets, and classification results from each subset are integrated. A boosting LDA is proposed in [48], where AdaBoost algorithm is adopted as the ensemble method. But the training error is often zero so that the boosting process cannot continue. To solve this problem, a cross-validation mechanism is used to weaken the LDA learner. [41] and [45] both use Bagging and random subspace methods for face recognition. In [41], each random subspace is spanned by $N_0 + N_1$ dimensions. The first N_0 dimensions are fixed as the N_0 top eigenvectors of the total covariance matrix. The remaining N_1 dimensions are randomly selected from the other $N - N_0 - 1$ eigenvectors. However, how to choose N_0 remains a big problem. In addition, the random subspace method has also been used with the Gaussian mixture model [44].

The remaining parts of this chapter is as follows: The problems in single- and multi-classifier face recognition algorithms are reviewed in Section 6.2. The proposed method is given in Section 6.3. The experiment results and discussion are given in Section 6.4 and the concluding remarks are drawn in the final section.

6.2 Problems in Single/Multiple-classifier System

6.2.1 Problems in Single-classifier System

It is well known that linear discriminant analysis normally requires the face images to follow a convex distribution, which may be approximately met in a small

database with limited variations in face images. However, the distribution dramatically becomes highly non-convex as the size of the database increases. To address this problem, two ways are usually adopted: (1) One is to model the complex distribution using nonlinear approaches, e.g., kernel methods [59,60]. (2) The other is to learn the complex distribution piecewise using a mixture of local linear model [61]. Since it is extremely inefficient to tune the nonlinear parameters in kernel method for the optimal performance and it is difficult to find a suitable kernel for a specific problem, the method of mixture of local linear model appears to be simpler, effective and computationally attractive. From the above derivation of Fisherface, it is easy to find that the obtained optimal transformation is a global and single projection matrix, and the extracted feature is incapable of generalizing all the image variations (This is a common problem for all the single classifiers mentioned above, e.g., null-space LDA [38]).

In Fisherface, although $\mathbf{S}_w$ is nonsingular in the PCA transformed low-dimensional space, $N - c$ dimensionality is still too high for the training set in many cases. When the training set is small (e.g., only two/three training samples available for each subject), $\mathbf{S}_w$ is not well estimated. A slight disturbance of noise on the training set will greatly change the inverse of $\mathbf{S}_w$. Therefore, the LDA classifier is often biased and numerically unstable. A natural way dealing with such a problem is to further reduce the dimensionality of the PCA transformed space to guarantee a more stable classifier. However, more discriminant information will certainly be discarded thereafter since it has been reported that the discriminant information actually exists in the whole $(N - 1)$ -dimensional range space of the total scatter matrix, $\mathbf{S}_t$ [46]. Therefore, there is a dilemma in the choice of the dimensionality of the PCA transformed space. In fact, the proper dimensionality depends on the training set. Figure 6.1 reports such a phenomenon. Figure 6.1 (a) reports that the optimal result does not appear at the 120th ($N - c = 40 \times 4 - 40$) dimension of PCA subspace when there are 4 training samples for each subject in ORL database. A similar case appears in Figure 6.1 (b) where the optimal PCA dimension is about

Figure 6.1: Recognition rate of Fisherface classifier with different dimension of PCA subspace.

60th instead of 240th ($40 \times 7 - 40$) when there are 7 training samples for each subject.

6.2.2 Problems in Multi-classifier System

It is well known that, in the construction of the classifier ensemble, the task of choosing the most suitable component classifiers or the kind of component classifiers is a key factor in improving the generalization ability of the ensemble classifier. The diversity between the component classifiers and the accuracy of each component classifier are known to be two important factors in affecting the generalization performance of ensemble classifiers [62, 63]. In practical design of ensemble classifier, if the component classifiers are too accurate, it is difficult to find diverse ones, which means that the errors made by these component classifiers should be highly correlated. The combination of these accurate but non-diverse classifiers often leads to very limited improvement. On the other hand, if the component classifiers are too inaccurate, although we can find diverse ones, the combination result may be worse than that of combining both more accurate and diverse component classifiers. This is because the combination result is dominated by too many inaccurate component classifiers, and will be wrong most of the time, leading to a poor generalization

result. These problems result in the dilemma between the accuracy and diversity of the component classifiers [64], i.e., the maximization of the diversity under the condition of obtaining a fairly accurate component classifiers is our objective.

Unfortunately, all the previous ensemble based face recognition approaches [41, 44, 45, 47, 48] create the ensemble classifier simply by a single ensemble method, e.g., bagging, or two ensemble methods separately, e.g., random subspace and bagging, without considering the diversities in selecting the component classifiers. Although in [41], a cross-validation approach is used to predetermine the fixed PCA dimension in random subspace LDA in order to achieve a better generalization by combining all the component classifiers, however, this technique turns out to be database-dependent and non-efficient. In this chapter, a framework of Ensemble Learning for Diversified Fisher Linear Discriminant (EL-DFLD) is proposed to improve the current Fisher linear discriminant based face recognition algorithms.

6.2.3 Our Solutions to the Problems in Single-classifier System

To overcome the dilemma in the choice of the dimensionality of the PCA transformed space in Fisherface method, Random Sub-feature LDA (RS-LDA) is proposed. Compared with the random subspace LDA in [41], the Random Sub-feature method in our paper is more complete from the perspective of "randomicity" because there's no fixed PCA subspace anymore in our method. Indeed, the random subspace LDA in [41] is only a quasi-random subspace method. To overcome the problem of non-convex distribution of face image data, AdaBoost-LDA (A-LDA) is proposed as one of the ensemble methods in this paper. AdaBoost-LDA shares the similar idea with the method of mixture of local linear model because of the following two points: (1) the use of Adaboost algorithm to generate a series of component LDA classifiers with each of which can generalize a local linear distribution of face images and a local linear solution can thus be obtained, and (2) a globally strong

solution through the combination of the multiple local solutions. Although it is believed to be unsuitable to apply AdaBoost algorithm to those stable base learner such as LDA because the effectiveness of Adaboost depends to a large extent on the the base learner's "instability", we will show that the LDA's "stability" will be sharply decreased by coupling Random Sub-feature LDA and AdaBoost-LDA. The reasons are as follows: Firstly, the component LDA classifiers constructed through RS-LDA in our method have a large variation in recognition performance due to the cancellation of the fixed PCA subspace in random sampling. Secondly, we adopt a new method for the computation of the training error in A-LDA to guarantee that the training error is nonzero.

6.2.4 Our Solutions to the Problems in Multi-classifier System

Compared with the previous ensemble-based face recognition methods [41, 44, 45, 47, 48], two major distinctions are given in this paper. Firstly, a new method is given to compute the diversities among the candidate component LDA classifiers. Secondly, considering the accuracy/diversity dilemma in ensemble learning, the classifier ensemble in EL-DFLD is composed of a set of diversified component LDA classifiers, which are selected intentionally based on the diversity computed in the last step. It is also verified in the experiment section that the coupling of Random Sub-feature LDA and AdaBoost-LDA will offer more suitable candidate component classifiers for ensemble methods in handling the dilemma of accuracy and diversity.

Therefore, four advantages arise from our coupling scheme. Firstly, we need not bother to find the dimensionality of fixed PCA subspace as in [41] so that our random sub-feature scheme is more efficient. Secondly, the algorithm of AdaBoost can also be applied to those stable and strong base learner such as LDA, which is very effective in face recognition. Thirdly, it can deal with the singularity, overfitting and bias problems that exist in FLD. Finally, such a coupling scheme will result in

more suitable component classifiers so as to increase the generalization performance of EL-DFLD. As an ensemble-based method, it can handle the accuracy/diversity dilemma in ensemble learning.

6.3 Ensemble Learning for Diversified Fisher Linear Discriminant

6.3.1 Random Sub-feature LDA

Although the dimension of image space is very high, only part of the full space contains the discriminative information. This subspace is spanned by all the eigenvectors of the total covariance matrix with nonzero eigenvalues. For the covariance matrix computed from N training samples, there are at most $N - 1$ eigenvectors with nonzero eigenvalues. Therefore, for Random Sub-feature LDA, we apply PCA to the face training set. All the eigenfaces with zero eigenvalues are removed, and $N - 1$ eigenfaces $\mathbf{U}_t = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{N-1}]$ are retained as candidates to construct the random subspaces.

Different from random subspace LDA adopted in [41], we do not add any constraint on the random sampling of the $\mathbf{U}_t$. We do not need to do a cross-validation to find the fixed PCA dimension as in [41] since these random sub-feature LDA classifiers are not required to be stable (In [41], the cross-validation is carried out to find the fixed PCA dimension for the stable classifier. On the contrary, our method does not need to do this since the diversity among the component classifiers is sought to construct the final classifier). Using such a random sampling method, we construct a lot of weak LDA classifiers. A more powerful classifier can be constructed by combining these component classifiers. A detailed description of RS-LDA is listed in Table 6.1.

Table 6.1: The algorithm of Random Sub-feature LDA**Algorithm:** RS-LDA

- 1. Input:** a set of training samples with labels $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, where $y_i \in Y = [1, 2, \dots, c]$; LDA algorithm; the number of sampling R_i for the given number of eigenvectors to be sampled, n_i .
- 2. Do:** Apply PCA to the face training set. All the eigenfaces with zero eigenvalues are removed, and $N - 1$ eigenfaces $\mathbf{U}_t = [u_1, u_2, \dots, u_{N-1}]$ are retained as candidates to construct the random subspaces.
- 3. Do for:** $i = 1, \dots, K$: Set the amount of the eigenvectors to be sampled, n_i . Randomly sample n_i eigenvectors for R_i times. Altogether $\sum_{i=1}^K R_i$ random subspaces are selected.
- 4. Do:** Perform LDA in the selected $\sum_{i=1}^K R_i$ random subspaces and $\sum_{i=1}^K R_i$ component LDA classifiers are obtained.
- 5. Output:** a set of $\sum_{i=1}^K R_i$ component component LDA classifiers.

6.3.2 AdaBoost-LDA

In this section, AdaBoost algorithm is applied to LDA base learner to create the A-LDA ensemble. As mentioned in Section 6.2, AdaBoost is generally applied to those unstable and weak base learner. Normally, the training error tends to be zero if LDA is used as base learner. To make AdaBoost applicable to LDA and overcome the zero training error problem, the subsampling technique [71] is used to generate a training subset: I_s , is sampled from the whole training set according to the weight distribution of training samples, where the number of samples in I_s is N_s . This technique guarantees that the sample with larger weight will be more likely selected. Compared with the cross-validation-like technique [48] where only the hardest examples are chosen, our technique is more in accordance to the principle of AdaBoost, since those samples with small weights will tend to be ignored completely in training stage in [48]. In addition, the cross-validation-like technique also easily results in overfitting problem. The detailed description of AdaBoost-LDA is listed in Table 6.2.

Table 6.2: The algorithm of AdaBoost-LDA**Algorithm:** AdaBoost-LDA**1. Input:** a set of training samples with labels $I = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, where $y_i \in Y = [1, 2, \dots, c]$; Fisherface algorithm, the number of cycles T .**2. Initialize:** the weight of samples: $w_i^1 = 1/N$, for all $i = 1, \dots, N$.**3. Do for** $t = 1, \dots, T$ (1) Sample a subset I_s from I according to the weight distribution of training samples, the number of samples in I_s is N_s .(2) Use Fisherface algorithm to train the weak-LDA classifier h_t on the selected weighted training sample set I_s .(3) Calculate the training error of h_t on I : $\epsilon_t = \sum_{i=1}^N w_i^t, y_i \neq h_t(\mathbf{x}_i)$.(4) Set weight of weak learner h_t : $\alpha_t = \frac{1}{2} \ln(\frac{1-\epsilon_t}{\epsilon_t})$.(5) Update training samples' weights: $w_i^{t+1} = \frac{w_i^t \exp\{-\alpha_t y_i h_t(\mathbf{x}_i)\}}{C_t}$ where C_t is a normalization constant, and $\sum_{i=1}^N w_i^{t+1} = 1$.**End****4. Output:** T component LDA classifiers with weights $\alpha_t, t = 1, 2, \dots, T$.

6.3.3 Coupling of AdaBoost-LDA and RS-LDA for Diversified FLD

Through AdaBoost-LDA and RS-LDA, two set of component classifiers can be constructed separately. However, what we want to show here is that we can also couple these two scheme to construct one set of component classifiers. Given a training set $\mathbf{I} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, a candidate component classifier set, $H = \{h_1, h_2, \dots\}$, can be constructed over $\mathbf{I}$ by using some component learning algorithms. Define a diversity-accuracy evaluation function, E , for judging the contribution of a candidate component classifier h to the ensemble classifier as:

$$E(h) = E(A(h), D(h)), h \in H \quad (6.1)$$

where E is a function of $A(h)$ and $D(h)$; $A(h)$ is a measure of the accuracy of h ; $D(h)$ is a measure of h 's diversity. In EL-DFLD, the h^* which can maximize the value of diversity-accuracy evaluation function is selected in each ensemble construction

iteration:

$$h^* = \arg \max_h E(A(h), D(h)), h \in H \quad (6.2)$$

Therefore, EL-DFLD works as follows in Table 6.3. Firstly, a sample set $\mathbf{I}_s$ is created from $\mathbf{I}$ by sub-sampling or weighted sampling techniques. Secondly, a candidate component classifier set, $H = \{h_1, h_2, \dots\}$, is trained by using LDA on $\mathbf{I}_s$. Thirdly, a diversity-accuracy (D-A) evaluation value, $E(h_i)$, is calculated for each $h_i \in H$ as $E(h_i) = (1 - w_{div}) * A(h_i) + w_{div} * D(h_i)$, where $A(h_i)$ is the training accuracy of h_i , $D(h_i)$ is the diversity of h_i and w_{div} is the weight of diversity, ranging from 0 to 1. Fourthly, the classifier in H with maximal evaluation value is selected as a component classifier of the final ensemble classifier. Afterwards, a new $\mathbf{I}_s$ is created and a new iteration begins until a given stop criterion, S , is satisfied. The final ensemble classifier, f , is a combination of selected component classifiers by using a fusion method F .

The function value of E is based on the values of $A(h)$ and $D(h)$. Normally, $A(h)$ has the scale from 0.5 to 1.0. However, the magnitude of $D(h)$ is constrained by the diversity of candidate component classifiers. That is, if the diversity over these classifiers is small, the diversity over the selected ones will not be large enough, although the diversity of the selected ones may be augmented through the ensemble learning algorithms, e.g. by training these classifiers on different subsets of samples in boosting. Hence, if the diversity of candidate component classifiers is constrained into a small scale, even if the selection criterion in Equation (6.2) is used, we cannot guarantee that the selected component classifiers with satisfied diversity level can be finally found. To get around this problem, a diversity augmentation method is always wanted. Fortunately, the diversity of the candidate component classifiers can be augmented by our coupling scheme. It has been shown in the previous section that the two-step LDA (Fisherface) with different PCA dimensions has different learning ability, see Figure 6.1. By adjusting the dimension of PCA subspace, Fisherface can demonstrate different learning ability. Hence, we use RS-LDA with different dimension of PCA subspace to obtain a set of component learning

algorithms with different learning abilities. The detailed EL-DFLD is described in Table 6.3. As in AdaBoost, the accuracy of a component classifier, h_t , is a weighted training accuracy:

$$A(h_t) = \sum_{i=1}^N \alpha_i \quad (6.3)$$

where $\alpha_i = 0$ if $y_i \neq h_t(\mathbf{x}_i)$ and $\alpha_i = w_i^t$ if $y_i = h_t(\mathbf{x}_i)$. w_i^t is the weight of training sample $\mathbf{x}_i$ when training classifier h_t . In the realization of proposed framework, we mainly aim to show the effectiveness of this framework, instead of finding the most suitable diversity measure. Hence, we just took one kind of diversity measure without any bias. However, our framework is open for other diversity measures. In EL-DFLD, the diversity of t -th classifier h_t is defined as follows: If $h_t(\mathbf{x}_i)$ is the prediction label of h_t on sample $\mathbf{x}_i$, and $f(\mathbf{x}_i)$ is the combined prediction label of all the existing classifiers, the diversity of h_t is defined as:

$$D(h_t) = \sum_{i=1}^N d_t(\mathbf{x}_i) \quad (6.4)$$

where $d_t(\mathbf{x}_i) = 0$ if $h_t(\mathbf{x}_i) = f(\mathbf{x}_i)$ and $d_t(\mathbf{x}_i) = 1$ if $h_t(\mathbf{x}_i) \neq f(\mathbf{x}_i)$.

6.4 Experiment Results and Discussion

EL-DFLD is used for face recognition and tested on ORL and Yale face database B. ORL database is used to evaluate the performance of EL-DFLD under conditions where the pose, face expression, face scale vary. Yale face database B is used to examine the performance when illumination varies extremely. We mainly design four types of experiments: the first type is to compare the performance of ensemble-based method with the consideration of D-A to those without considering D-A, e.g., RS-LDA with D-A vs. RS-LDA without D-A. The second type is to examine whether the recognition performance can be improved by applying AdaBoost method to the LDA base learner. The third type is to compare the performance of EL-DFLD

Table 6.3: The algorithm of Ensemble Learning for Diversified Fisher Discriminant**Algorithm:** EL-DFLD

1. Input: a set of training samples with labels $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, where $y_i \in Y = [1, 2, \dots, c]$; Fisherface algorithm; the number of cycles T, R .

2. Initialize: the weight of samples: $w_i^1 = 1/N$, for all $i = 1, \dots, N$.

3. Do for $t = 1, \dots, T$

(1) Sample a subset I_s from I according to the weight distribution of training samples, the number of samples in I_s is N_s .

(2) Perform PCA on I_s and obtain the $N_s - 1$ eigenvectors corresponding to the top $N_s - 1$ eigenvalues.

(3) Perform RS-LDA on the $N_s - 1$ eigenvectors obtained in (2). Assume that the amount of the eigenvectors to be sampled is set to be $n_j, j = 1, \dots, K$. For the specific n_j , randomly sample R_j times. Thus we obtain $n_{rs} = \sum_{j=1}^K R_j$ LDA component classifiers. Represent these component classifiers as $H = h_t^1, h_t^2, \dots, h_t^{n_{rs}}$.

(4) Calculate the accuracy of h_t^k : $A(h_t^k) = 1 - \sum_{i=1}^N w_i^t [h_t^k(\mathbf{x}_i) \neq y_i]$, $h_t^k \in H$.

(5) Calculate the diversity of h_t^k : $D(h_t^k) = \sum_{i=1}^N d_t(\mathbf{x}_i)$, $h_t^k \in H$.

(6) Calculate the diversity-accuracy evaluation value of h_t^k : $E(h_t^k) = (1 - w_{div}) * A(h_t^k) + w_{div} * D(h_t^k)$, $h_t^k \in H$.

(7) Select component classifier for the $E_n - DFLD$: $h_t = \arg \max_{h_t^k} E(h_t^k)$ and $A(h_t^k) > 50\%$, $h_t^k \in H$.

(8) Update the weights of training samples: $w_i^{t+1} = \frac{w_i^t (\frac{1-A(h_t)}{h_t})^{1-[h_t(\mathbf{x}_i) \neq y_i]}}{C_t}$, where C_t is a normalization constant, and $\sum_{i=1}^N w_i^{t+1} = 1$.

End

4. Output: $f(\mathbf{x}) = \arg \max_{y \in Y} \sum_{t=1}^T (\log \frac{A(h_t)}{1-A(h_t)}) [h_t(\mathbf{x}) = y]$

with RS-LDA and A-LDA. The fourth is to compare EL-DFLD with the previous LDA-based recognition methods.

6.4.1 Performance of EL-DFLD with/without D-A Consideration

To check the performance of our algorithm in dealing with the diversity/accuracy (D-A) dilemma, the following three tests are carried out: (1) Test of RS-LDA with the consideration of D-A dilemma, RS-LDA with D-A in short; (2) test of RS-LDA

Figure 6.2: The effect of our algorithm in dealing with D-A. (a): Comparison on ORL database; (b): Comparison on YaleB.

without considering D-A dilemma and no fixed PCA dimension in random sampling, (3) test of RS-LDA without considering D-A dilemma but with fixed PCA dimension in random sampling as in [41]. In RS-LDA, the amount of PCA dimension is set as $(N-1)/3$, $(N-1)/2$ and $2(N-1)/3$, where N is the total number of training samples. For each one, 40 random samplings are performed. Majority voting is used for the final decision. We compare the three test results with Fisherface method. Figure 6.2 demonstrates the comparison of performance on ORL and YaleB face databases. It can be seen that by dealing with the D-A dilemma, we may improve the recognition performance by 2 – 4% compared with its counterparts without considering D-A dilemma. The best performance is achieved when the w_{div} is approximately 0.4 for the two databases.

6.4.2 Performance of AdaBoost-LDA

This experiment aims to argue that AdaBoost can also be applied to those stable classifiers such as LDA by making LDA an unstable classifier. We have managed to make the training error of LDA to be nonzero. Figure 6.3 (a) and (b) report the comparison between A-LDA and other methods such as Fisherface and N-LDA on

ORL and YaleB face databases. A-LDA has an improvement of 2% over Fisherface and 5% over N-LDA on ORL face database. On Yaleface database, the improvement is much larger, especially when the number of training samples is small, e.g., it has an improvement of 5% over Fisherface and 8% over N-LDA when there are three training samples for each subject. This is in accordance with our analysis in Section 6.2, since the face images in YaleB face database has a large illumination variation, which results in a non-convex distribution of face image data. Since the use of Adaboost algorithm will generate a series of component LDA classifiers, and each of these LDAs can generalize a local linear distribution of face images and a local linear solution can thus be obtained. By combining these multiple local solutions, a globally strong solution can be obtained.

6.4.3 Performance by Coupling RS-LDA and AdaBoost-LDA

Through this experiment, we want to show that the result obtained from the EL-DFLD coupling scheme is better than both RS-LDA and A-LDA. Figure 6.3 (a) and (b) give the comparison with RS-LDA and A-LDA on ORL and YaleB face database. EL-DFLD outperforms RS-LDA with D-A by 5 – 6% and A-LDA by 7 – 8% on ORL face database. A larger improvement appears when performed on YaleB face database. This is consistent with our analysis that the combination of A-LDA and RS-LDA can achieve better generalization ability and produce larger diversity among the component classifiers.

6.4.4 Performance Comparison with Previous Methods

This experiment is to compare the performance of EL-DFLD with the previous LDA based recognition methods, e.g., Fisherface, N-LDA, 2DFDA. We test the recognition performance with different training numbers. k ($3 \leq k \leq 9$) images of each subject are randomly selected for training and the remaining $M - k$ (for ORL, M is 10; for YaleB, M equals 64) images of each subject for testing. For each k , 30

Figure 6.3: The comparison of performance of EL-DFLD with the other LDA-based methods in recognition. (a): Comparison on ORL database; (b): Comparison on YaleB.

runs are performed with different random partition between training set and testing set. For each run, EL-DFLD method is performed by training the selected fixed samples and testing on the left images. Figure 6.3 shows the average recognition rate. The best performance is achieved when the w_{div} is about 0.4. It can be seen that EL-DFLD outperforms 2DFDA by up to 2%-4%, surpasses RS-LDA by up to 6%, is better than A-LDA by up to 12%, and higher than Fisherface by up to 20%. From Figure 6.3, it can be seen that RS-LDA with D-A is better than A-LDA.

6.5 A comparative study of the four proposed linear methods

In this section, as required by one examiner, the performance comparisons of the four proposed linear methods are given with regards to accuracy and computational requirements. The four proposed linear methods are Bilateral 2DPCA (B2DPCA), Bilateral 2DFDA (B2DFDA), Modified FDA (MFDA) and Ensemble learning for Diversified FLD (EnEn-DFLD).

Table 6.4: Comparison of recognition performance on ORL database: the first row indicates the number of training sample for each person and the first column indicates the four proposed methods.

	1	2	3	4	5
B2DPCA	72.8%	85.1%	90.3%	93.5%	95.4%
B2DFDA	——	87.7%	92.7%	95.2%	96.8%
MFDA	——	84.2%	90.6%	94.0%	96.1%
En-DFLD	——	85.5%	94.5%	96.6%	97.8%

Table 6.5: Comparison of recognition performance on YaleB database: the first row indicates the number of training sample for each person and the first column indicates the four proposed methods.

	1	2	3	4	5
B2DPCA	36.2%	40.5%	45.7%	50.6%	54.9%
B2DFDA	——	60.5%	75.4%	78.6%	82.8%
MFDA	——	56.8%	73.9%	77.1%	82.2%
En-DFLD	——	55.9%	76.7%	81.8%	85.0%

6.5.1 Comparison of recognition performance on ORL and YaleB face databases

From the above tables, we may find that En-DFLD is the best among the four algorithms with regards to accuracy. The performance of the four algorithms on ORL database is better than that on YaleB database since YaleB face database has larger illumination variations than ORL. Considering the accuracy performance in the above tables and the computational complexity in Table 6.6 (in the following part for comparison of computational requirements), we may make some recommendations as requested by Examiner No.2. If there are multiple training samples for each subject and the computational complexity is an important factor, MFDA and B2DFDA are recommended. If there is only one training sample for each subject and no large illumination variations exist, B2DPCA is recommended. If only two or three training samples for each subject are available, B2DFDA is recommended.

Table 6.6: Comparison of computational requirements, i.e., CPU Time (s), for feature extraction on ORL face database (CPU: Pentium IV 2.4GHz, RAM: 512Mb): the first row indicates the number of training sample for each person and the first column indicates the methods

	2	3	4	5
LDA	1.125s	2.391s	4.281s	6.75s
B2DPCA	0.343s	0.797s	1.359s	2.215s
B2DFDA	0.547s	0.922s	1.547s	2.312s
MFDA	0.706s	1.256s	1.662s	2.871s
En-DFLD	5.377s	12.812s	29.246s	40.36s

6.5.2   Comparison of computational requirement

As required by Examiner No.2, a comparison study of computational requirement of the four algorithms is given in this section. To compare the computational requirements with PCA and LDA, we have listed the time in feature extraction using B2DPCA, B2DFDA, MFDA and En-DFLD. Since LDA uses more time than PCA in feature extraction, we only compare our algorithms with LDA (Fisherface). Note that the time is calculated using Matlab 7.0. From the comparison, we may find that the 2D based algorithms are more efficient than conventional vector based algorithm in feature extraction. En-DFLD is the most expensive one among the four algorithms since it will use the boosting process to select the most discriminative features.

6.6   Concluding Remarks

In this chapter, a framework of Ensemble Learning for Diversified Fisher Linear Discriminant (EL-DFLD) is proposed for face recognition. In EL-DFLD, an AdaBoost-LDA (A-LDA) and a Random Sub-feature LDA (RS-LDA) schemes are proposed and coupled to construct the total LDA classifier ensemble. By dealing with the diversity/accuracy dilemma through such a coupling scheme, recognition accuracy can be significantly improved. Experiments on two public face databases verify

the superiority of the proposed EL-DFLD over the state-of-the-art algorithms in recognition accuracy.

Chapter 7

Conclusion and Recommendations

7.1 Conclusion

In this thesis, we have mainly addressed the Small Sample Size problem in face recognition via four subspace analysis algorithms.

Firstly, a Generalized 2D Principal Component Analysis is given to overcome the limitations of 2DPCA. The theoretical analysis is given to explain why 2DPCA is better than PCA. It is revealed that 2DPCA performed on the 2D images is essentially a clustered PCA performed on all the rows of all the images if each row is viewed as a computational unit. In addition, 2DPCA needs much more coefficients to represent an image than PCA. Correspondingly, a Bilateral 2DPCA is proposed to overcome this problem. Furthermore, a Kernel 2DPCA (K2DPCA) scheme is introduced to improve the generalization ability and K2DPCA has been shown to be essentially the clustered KPCA method performed on image rows/columns. Experiment results on ORL, Yale and UMIST databases show that B2DPCA is more efficient in image representation and recognition. K2DPCA outperforms PCA, 2DPCA and KPCA by up to 5%.

Secondly, the 2D Fisher Discriminant Analysis is proposed to deal with the Small

Sample Size problem in Fisher Linear Discriminant. 2DFDA is based on 2D image matrices rather than 1D column vectors so that the image matrix does not need to be transformed into a long vector before feature extraction. The essence of 2DFDA is revealed that the 2DFDA performed on the image matrices is a clustered FLD performed on the rows/columns of all the training samples. The advantage arising in this way is that the between-class and within-class scatter matrices constructed in 2DFDA are generally both of full rank. This framework has shown much better performance in dealing with SSS problem. The experiments on ORL, YaleB and UMIST databases validated that 2DFDA outperforms FLD, D-LDA and N-LDA by up to 10% in recognition rate when there are very few training samples and K2DFDA is better than Kernel Fisherface in recognition accuracy by up to 4%.

Next, a framework of low-dimensional Fisher discriminant subspace analysis is proposed. In this framework, it is shown that the null space of the total covariance matrix contributes little to discrimination. Therefore, a low-dimensional feature space is developed by removing the null space of the total covariance matrix. Two different algorithms are then proposed in this low-dimensional feature space. Unified Linear Discriminant Analysis extracts discriminant information from three subspaces of the range space of S_t . Modified Linear Discriminant Analysis can avoid the numerical problem in conventional LDA by adopting a Deduction-form Fisher criterion in place of the Quotient-form one. Furthermore, the Kernel version of MLDA is also developed to extract the nonlinear discriminant vectors. Experimental results on a large combined database have demonstrated that the proposed ULDA and MLDA can both achieve better performance than the current LDA-based algorithms by up to 4%, and KMFDA can outperform Kernel Fisherface by up to 3% in recognition accuracy.

Finally, a framework of Ensemble Learning for Diversified Fisher Linear Discriminant is proposed to deal with the existing problems in the LDA-based single-classifier and multiple-classifier face recognition algorithms. In this framework, an AdaBoost-LDA and a Random Sub-feature LDA schemes are incorporated to construct the

total weak-LDA classifier ensemble. By such a coupling scheme, we can deal with the diversity/accuracy dilemma in ensemble method, recognition accuracy can be significantly improved. Experiments on ORL and YaleB databases have shown that Ensemble Learning for Diversified Fisher Linear Discriminant can outperform the current LDA-based methods. Specifically, it can outperform 2DFDA by up to 2% to 3% in accuracy.

The first two algorithms adopt the matrix-based data representation model while the last two algorithms use the vector-based data representation model. All the four algorithms can outperform the previous subspace analysis method when there are very limited training samples, and the last three algorithms are to some extent tolerable to the image variations caused by pose, illumination and expression.

7.2 Recommendations for Future Research

We conclude this thesis by identifying some directions for future research work in this area.

(1) Ensemble Kernel Fisherface

Kernel Fisher Discriminant Analysis (Kernel Fisherface) has aroused considerable interest in the fields of pattern recognition and machine learning. However, Kernel Fisherface always encounters the ill-posed problem in its real-world applications, i.e., the Small Sample Size problem. Although several algorithms have been proposed to overcome this problem, unfortunately, all of these methods discard some discriminant information due to the limitation of these algorithms.

Recently, the essence of Kernel Fisherface has been revealed, i.e., Kernel Fisherface is equivalent to Kernel PCA plus Fisherface. Based on the two-step Kernel Fisherface framework, a framework of Ensemble Kernel Fisherface (EnKF) is proposed to deal with the Small Sample Size problem in kernel Fisherface

(kernel Fisher discriminant analysis). In many cases, kernel Fisherface classifier is overfitted to the training set and discard some useful discriminative information. By representing kernel Fisherface method in the new two-step KFD framework, we give novel insights into the two reasons that arouse the overfitting problem. The first one is the non-representative training samples. The dimension of the reserved KPCA subspace is still too high. The second one is that the covariance matrices in the KPCA-transformed space are still singular. $\mathbf{S}_w^\Phi$ is not well estimated. A slight disturbance of noise on the training set will greatly change the inverse of $\mathbf{S}_w^\Phi$. A Boosting-KFD (B-KFD) and a Random Sub-feature KFD (RS-KFD) schemes are proposed respectively to reduce the effect of these two intrinsic factors. EnKF is constructed by coupling B-KFD and RS-KFD to produce the total weak-KFD classifier ensemble.

(2) An Algorithm Dealing with Diversity/Accuracy Dilemma in Ensemble Kernel Fisherface

By analyzing the overfitting problem in the two-step Kernel Fisherface approach, Ensemble Kernel Fisherface (EnKF) will be proposed as a future research work in this thesis. However, it is well-know that there exists the diversity/accuracy dilemma in ensemble methods. Therefore, a new algorithm that handles the diversity/accuracy dilemma can be embedded in the EnKF framework to handle the diversity/accuracy dilemma of ensemble methods. In the new algorithm, a criterion involving both diversity and accuracy can be defined, and each component classifier of the final ensemble classifier can be selected by maximizing this criterion. Moreover, a new diversity augmentation methodology can be proposed by purposely expanding the scale of learning ability over candidate component classifiers. The diversity augmentation algorithm is proposed by combining a random subspace method and the diversity value among the component classifiers can be increased by (1) purposely adjusting the kernel parameter, σ and (2) using random subspace method to randomly select the kernel PCA subspace. Different from the con-

ventional diversity augmentation method in which all of these methods only focused on varying training data to create diversity, such as using different feature subsets, or using different weights for training data, or bootstrapping a part of whole training set. The proposed method can also augment diversity through adjusting the learning ability of the component classifiers. This method can effectively improve the diversity and give more suitable candidate component classifiers to EnKF resulting in a better generalization performance. A realization of EnKF can also be developed by using B-LDA, RS-LDA, or Bagging-LDA as component classifiers.

(3) Face Recognition Based on Multi-view Camera System

It is a great challenge for face recognition algorithms using only 2D grey-level images to recognize faces with significant pose and illumination changes. However, if multi-camera imaging system is used for face recognition, a pose- and illumination-invariant face recognition system can be constructed easily. This is because the dense depth (disparity) images can be accurately estimated via the multi-view epipolar geometry, and the depth image is invariant to illumination variations. Since it has been shown that the combination of the 2D grey-level and 3D disparity images can improve face recognition significantly. Thus, the depth and grey-level images can be combined together to improve the recognition performance. In addition, it is also pose-invariant since we can get the multi-view face images for the same subject, and thus use the multi-view face images for training purposes. If we use wide-baseline multi-camera system, it is difficult to obtain the dense disparity images. We will adopt a view-based and modular kernel fisherface for face recognition. Another method for dealing with the wide-baseline multi-camera system is to use the affine-invariant local-area correspondence algorithm to find the affine-invariant local areas of the same subject, and then find the corresponding areas of the test image. By applying component-based kernel fisherface method to these areas and combine these results using a fusion scheme, it is expected to

improve the recognition performance.

(4) 2D/3D Face Recognition Based on Video Sequences

With the development of the current hardware technology, the contents of digital libraries consist of increasingly large amounts of digital media and, in particular, pictorial information characteristic of video data can be easily obtained and stored. Therefore, this makes it possible to investigate the 2D/3D face recognition algorithms from video sequences. By combining the conventional algorithms in 3D computer vision, e.g., structure from motion, shape from shading and photometric stereo etc, the 3D face geometry can be estimated. By using image-based rendering techniques, multiple synthesized images could be obtained for training. This will greatly enhance the size of training set. In addition, by further investigating the discriminant information from the available video sequences (2D images), the combination of this two types of training data will significantly improve the performance of recognition. With enough training samples, machine learning techniques such SVM and AdaBoost can further be applied to get a better generalization ability for the classifiers.

Author's Publications

- [1] Hui Kong, Lei Wang, Eam Khwang Teoh, Jian-gang Wang, Ronda Venkateswarlu "Generalized 2D Principal Component Analysis for Face Image Representation and Recognition", *Neural Networks*, 18(5-6): 585-594, 2005
- [2] Hui Kong, Eam Khwang Teoh, Matthew Turk, and Jian-Gang Wang "Coplanar-light Sweep Surface Supported Uncalibrated Photometric Stereo", to appear in *Journal of Optical Society of America, A*, Dec 2006.
- [3] Hui Kong, Xuchun Li, Eam Khwang Teoh and Matthew Turk, "Ensemble Learning for Subspace Analysis with Diversity", Under revision of *IEEE Transactions on Image Processing*, Jul 2005
- [4] Hui Kong and Eam Khwang Teoh, "Margin-Maximized Discriminant Analysis for Multi-shot Based Object Recognition", submitted to *Neural Computing*, Jul 2006
- [5] Hui Kong, Eam Khwang Teoh, and Pengfei Xu, "Margin-Maximized Discriminant Analysis for Multi-shot Based Object Recognition," in Proceedings of *International Symposium on Vision Computing (ISVC'06)*, USA, 06-08 Nov 2006
- [6] Hui Kong, Pengfei Xu, and Eam Khwang Teoh, "Binocular Uncalibrated Photometric Stereo", in Proceedings of *International Symposium on Vision Computing (ISVC'06)*, USA, 06-08 Nov 2006

- [7] Hui Kong, Eam Khwang Teoh, "Coupling Adaboost and Random Subspace for Diversified Fisher Linear Discriminant", in Proceedings of *International Conference on Control, Automation, Robotics and Vision (ICARCV'06)*, Singapore, 5-8 Dec 2006
- [8] Hui Kong and Eam Khwang Teoh, "Ensemble LDA for Face Recognition", in Proceedings of *International Symposium on Vision Computing (ISVC'05)*, USA, 05-07 Dec 2005
- [9] Hui Kong and Eam Khwang Teoh, "Discriminant Analysis of Low-dimensional Space for Face Recognition with Small Number of Training Samples" in Proceedings of *British Machine Vision Conference (BMVC'05)*, Oxford, UK, 05-08 Sep 2005
- [10] Hui Kong and Eam Khwang Teoh, "Generalized 2D Fisher Discriminant Analysis," in Proceedings of *British Machine Vision Conference (BMVC'05)*, Oxford, UK, 05-08 Sep 2005
- [11] Hui Kong, Xuchun Li, Lei Wang, Eam Khwang Teoh, Jian-gang Wang and Ronda Venkateswarlu, "Generalized 2D Principal Component Analysis," in Proceedings of *International Joint Conference on Neural Networks (IJCNN'05)*, Montreal, Canada, 31 Jul - 04 Aug 2005
- [12] Hui Kong, Lei Wang, Eam Khwang Teoh, Jian-gang Wang and Ronda Venkateswarlu, "A Framework of 2D Fisher Discriminant Analysis: Application to Face Recognition with Small Number of Training Samples," in Proceedings of *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol.2, pp. 1083-1088, San Diego, USA, 20-25 Jun 2005
- [13] Hui Kong, Eam Khwang Teoh, Jian-gang Wang and Ronda Venkateswarlu, "Two Dimensional Fisher Discriminant Analysis: Forget about Small Sample Size Problem," in Proceedings of *IEEE International Conference on Acoustics,*

Speech, and Signal Processing (ICASSP'05), Vol.2, pp.761-764, USA, 19-23 Mar 2005

- [14] Hui Kong, Eam Khwang Teoh, Jian-gang Wang and Ronda Venkateswarlu, "Coplanar Light Sweep-Surface Supported Uncalibrated Photometric Stereo," in Proceedings of *International Conf. on Pattern Recognition* (ICPR'04), Vol.4, pp.92-95, Cambridge, UK, 23-26 Aug 2004
- [15] Jian-gang Wang, Hui Kong and Ronda Venkateswarlu, "Improving Face Recognition Performance by Combining Color and Depth Fisherfaces," in Proceedings of *Asian Conf on Computer Vision* (ACCV'04), pp.126-131, Jeju Island, Korea, 27-30 Jan 2004
- [16] Hui Kong, Eam Khwang Teoh, Jian-gang Wang and Ronda Venkateswarlu, "Eigen-coefficients Estimation of Occluded Face Using selection and clustering," in Proceedings of *Asian Conf on Computer Vision* (ACCV'04), pp.150-155, Jeju Island, Korea, 27-30 Jan 2004
- [17] Hui Kong and Eam Khwang Teoh, "Generic Face Model Adaptation Based on Non-rigid Feature Points Correspondence," in Proceedings of IEEE Pacific-Rim Conference On Multimedia (PCM'03), Singapore, 15-18 Dec 2003

Bibliography

- [1] P.J. Phillips, H. Wechsler, J. Huang and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms", *Image and Vision Computing*, Vol.16, no.5, pp.295-306, 1998
- [2] P.J. Phillips, H. Moon, S. Rizvi and P. Rauss, "The FERET evaluation methodology for face recognition algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, no.10, pp.1090-1104, Oct 2000
- [3] P.J. Phillips, P.J. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi and J.M. Bone, "Face recognition vendor test 2002: Evaluation report," NISTIR 6965, 2003. Available online at [http : //www.frvt.org](http://www.frvt.org).
- [4] Face Recognition Vendor Tests, [http : //www.frvt.org/](http://www.frvt.org/).
- [5] Face Recognition Grand Challenge, [http : //www.frvt.org/FRGC/](http://www.frvt.org/FRGC/).
- [6] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre, " XM2VTSDB: The Extended M2VTS Database," in Proceedings of *International Conference on Audio- and Video-Based Person Authentication (AVBPA)*, Washington DC, pp.72-77, 1999
- [7] P.J. Phillips, R.M. McCabe and R. Chellappa, "Biometric image processing and recognition," in Proceedings of *European Signal Processing Conference*, 1998.

- [8] R. Chellappa, C. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," in *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705-740, 1995.
- [9] J. Yang, D. Zhang, A. F. Frangi and J. Yang, "Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131-137, 2004
- [10] L. Sirovich and M. Kirby, "Low-Dimensional Procedure for the Characterization of Human Faces," *Journal of Optical Society of America, A*, vol. 4, no. 3, pp. 519-524, 1987
- [11] M. Kirby and L. Sirovich, "Application of the KL Procedure for the Characterization of Human Faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103-108, 1990
- [12] M. Turk and A. Pentland, "Eigenfaces for Recognition," 3, 71-86 *Journal of Cognitive Neuroscience*, 1991
- [13] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," in *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 586-591, 3-6 Jun 1991
- [14] A. Pentland, B. Moghaddam and Thad Starner, "View-based and Modular Eigenspaces for Face Recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.84-91, Jun 1994
- [15] B. Heisele, P. Ho, J. Wu and T. Poggio, "Face recognition: component-based versus global approaches", *Computer Vision and Image Understanding*, 2003.
- [16] H. Murase and S. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance," *International Journal on Computer Vision*, vol. 14, No. 1, pp. 5-24, 1995

- [17] A. Shashua, "Geometry and Photometry in 3D Visual Recognition," Ph.D Thesis, MIT, 1994
- [18] P. Hallinan, "A Low-Dimensional Representation of Human Faces for Arbitrary Lighting Conditions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 995-999, Jun 1994
- [19] Y. Moses, Y. Adini, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, No. 7, pp. 721-732, 1997
- [20] R. Epstein, P. Hallinan and A. L. Yuille, " 5 ± 2 Eigenimages Suffice: An Empirical Investigation of Low-Dimensional Lighting Models," in *Proceedings of IEEE Workshop on Physics-Based Modeling in Computer Vision*, pp.108-116, 1995
- [21] L. Zhao and Y. Yang, "Theoretical Analysis of Illumination in PCA-Based Vision Systems," *Pattern Recognition*, vol. 32, pp. 547-564, 1999
- [22] R. Ramamoorthi, "Analytic PCA Construction for Theoretical Analysis of Lighting Variability in Images of a Lambertian Object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 10, pp. 1322-1333, Oct 2002
- [23] H. Bischof, H. Wildenauer, and A. Leonardis, "Illumination insensitive recognition using eigenimages", *Computer Vision and Image Understanding*, 95, pp. 86-104, 2004
- [24] X. Liu, T. Chen and B.V.K. Vijaya Kumar, Face Authentication for Multiple Subjects Using Eigenflow. *Pattern Recognition*, special issue on Biometric, vol. 36, no. 2, pp. 313-328, Feb 2003
- [25] S.Z. Li and J. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. on Neural Networks*, vol. 10, pp. 439-443, Mar 1999

- [26] R.A. Fisher, "The Use of Multiple Measures in Taxonomic Problems," *Annual Eugenics*, vol. 7, pp. 179-188, 1936
- [27] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of Optical Society of America, A* 14, pp. 1724-1733, 1997
- [28] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831-836, 1996
- [29] D.L. Swets and J. Weng, "Discriminant analysis and eigenspace partition tree for face and object recognition from views," In Proceedings of *International Conference on Automatic Face and Gesture Recognition*, pp. 192-197, 1996
- [30] P. N. Belhumeur and J. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997
- [31] W. Zhao, "Discriminant Component Analysis for Face Recognition," in Proceedings of *Int. Conf. on Pattern Recognition*, vol. 2, pp. 818-821, 2000
- [32] W. Zhao, "Robust Image Based 3D Face Recognition," Ph.D. dissertation. University of Maryland, College Park, MD 2000
- [33] W. Zhao, R. Chellappa and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," in Proceedings of *International Conference on Automatic Face and Gesture Recognition*, pp. 336-341, 1998
- [34] W. Zhao, R. Chellappa and P.J. Phillips, "Subspace linear discriminant analysis for face recognition," Tech. rep. CAR-TR-914, Center for Automation Research, University of Maryland, College Park, MD, 1999
- [35] W. Zhao and R. Chellappa, "Illumination-insensitive face recognition using symmetric shape-from-shading," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 286-293, 2000

- [36] W. Zhao and R. Chellappa, "SFS Based View synthesis for robust face recognition," in Proceedings of *International Conference on Automatic Face and Gesture Recognition*, pp. 285-292, 2000
- [37] C. Liu and H. Wechsler, "A shape- and texture-based enhanced fisher classifier for face recognition," *IEEE Trans. on Image Processing*, vol. 10, no. 4, pp. 598C608, 2001
- [38] L. Chen, H. Liao, M. Ko, J. Lin, and G. 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, 2000
- [39] 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, 2001
- [40] R. Huang, Q.S. Liu, H.Q. Lu and S.D. Ma, "Solving the Small Sample Size Problem of LDA," in Proceedings of *Internatinal Conference on Pattern Recognition*, vol.3, pp.29-32, 2002
- [41] X. Wang and X. Tang, "Random sampling LDA for face recognition," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 259-265, 2004
- [42] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832C844, Aug 1998
- [43] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 02, no. 2, pp. 564-569, Jun 2004
- [44] X. Wang and X. Tang, "Subspace Analysis Using Random Mixture Models," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 574-580, Jun 2005

- [45] N.V. Chawla and K.W. Bowyer, "Random Subspaces and Subsampling for 2-D Face Recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 582-589, Jun 2005
- [46] H. Kong, X. Li, E.K. Teoh, J.G. Wang and C. Kambhamettu, "Discriminant Analysis of Low-dimensional Space for Face Recognition with Small Number of Training Samples," in *Proceedings of British Machine Vision Conference*, 2005
- [47] X. Lu and A.K. Jain, "Resampling for Face Recognition," in *Proceedings of International Conference On Audio- And Video-Based Biometric Person Authentication*, pp. 869-877, 2003
- [48] J.W. Lu, K.N. Plataniotis and A.N. Venetsanopoulos, "Boosting Linear Discriminant Analysis For Face Recognition," in *Proceedings of IEEE International Conference on Image Processing*, 2003
- [49] H. Cevikalp, M. Neamtu, M. Wilkes and A. Barkana, "Discriminative Common Vectors for Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 4-13, Jan 2005
- [50] J. Ye, R. Janardan, C.H. Park and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 982-994, 2004
- [51] S. Dudoit, J. Fridlyand and T.P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77C87, 2002
- [52] S. Raudys and R.P.W. Duin, "On expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix," *Pattern Recognition Letter*, vol. 19, no. 5-6, pp. 385-392, 1998

- [53] M. Skurichina and R.P.W. Duin, "Stabilizing classifiers for very small sample size," in Proceedings of *International Conference on Pattern Recognition*, vol. 2, pp. 891-896, 1996
- [54] D.Q. Dai and P.C. Yuen, "Regularized discriminant analysis and its application to face recognition," *Pattern Recognition*, vol. 36, no. 3, pp. 845-8, 2003
- [55] J. Lu and K.N. Plataniotis and A.N. Venetsanopoulos, "Regularized Discriminant Analysis For the Small Sample Size Problem in Face Recognition," *Pattern Recognition Letter*, vol. 24, no. 16, pp. 3079C3087, 2003
- [56] K. Fukunnaga, *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, Boston 1990
- [57] R.A. Johnson and D.W. Wichern, "Applied Multivariate Statistical Analysis," fourth edition, Prentice-Hall, Inc, 1998
- [58] 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
- [59] M. H. Yang, "Kernel Eigenface vs. Kernel Fisherface: Face Recognition Using Kernel Methods," in Proceedings of *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 215-220, May 2002
- [60] Q. Liu, R. Huang, H. Lu and S. Ma, "Face Recognition Using Kernel Based Fisher Discriminant Analysis," in Proceedings of *IEEE International Conference on Automatic Face and Gesture Recognition*, pp.197-201, 2002
- [61] T.K. Kim and J. Kittler, "Locally Linear Discriminant Analysis for Multimodally Distributed Classes for Face Recognition with a Single Model Image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 318-327, 2005

- [62] L.I. Kuncheva and C.J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181-207, 2003
- [63] T. Windeatt, "Diversity measures for multiple classifier system analysis and design," *Information Fusion*, vol. 6, pp. 181-207, 2003
- [64] T.G. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization," *Machine Learning*, vol. 40, no. 2, pp. 21-36, 2005
- [65] G. Baudat and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach", *Neural Computation*, vol. 12, no. 1, pp. 2385C2404, 2000
- [66] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms", *IEEE Transactions on Neural Networks*, vol. 14, No. 1, pp. 117-126, Jan 2003
- [67] J. Ye, R. Janardan and Q. Li, "Two-Dimensional Linear Discriminant Analysis," in Proceedings of *Neural Information Processing Systems (NIPS)*, Dec 2004
- [68] H. Kong, X. Li, L. Wang, E.K. Teoh, J.G. Wang and V. Ronda, "Generalized 2D Principal Component Analysis," in Proceedings of *International Joint Conference on Neural Networks (IJCNN05)*, Montreal, Canada, 31 Jul - 04 Aug 2005
- [69] Hui Kong, Lei Wang, Eam Khwang Teoh, Jian-gang Wang, Ronda Venkateswarlu "Generalized 2D Principal Component Analysis for Face Image Representation and Recognition", *Neural Networks*, 18(5-6): 585-594, 2005
- [70] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643-660, 2001

- [71] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1999
- [72] B.T. Phong, "Illumination for Computer Generated Pictures," *Communications of the ACM*, vol. 18, no. 6, pp. 311-317, 1975
- [73] W. Zhou and C. Kambhamettu, "Estimation of Illuminant Direction and Intensity of Multiple Light Source", in *Proceedings of European Conference on Computer Vision*, vol. 4, pp. 206-220, Copenhagen, Denmark, May 27-Jun 2, 2002
- [74] K. Torrance and E. Sparrow, "Theory for Off-Specular Reflection From Roughened Surfaces," *Journal of the Optical Society of America*, vol. 57, no. 9, pp. 1105-1114, Sep 1967
- [75] A.S. Georgiades, "Incorporating the Torrance and Sparrow Model of Reflectance in Uncalibrated Photometric Stereo," in *Proceedings of International Conference on Computer Vision*, pp. 816-823, Nice, France, Oct 2003
- [76] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 696-710, 1997
- [77] G. D. Guo, S. Z. Li, and K. L. Chan, "Face Recognition by Support Vector Machines," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pp.196-201, Grenoble, France, 2000
- [78] M.S. Bartlett, J.R. Movellan and T.J. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450-64, 2002
- [79] J. Draper, K. Baek, M.S. Bartlett and J.R. Beveridge, "Recognizing faces with PCA and ICA," *Computer vision and image understanding 91: Special issue on Face Recognition*, pp. 115-137, 2003

- [80] S.H. Lin, S.Y. Kung and L.J. Lin, "Face recognition/detection by probabilistic decision based neural network," *IEEE Transactions on Neural Networks*, vol. 8, pp. 114C132, 1997
- [81] C. Liu and H. Wechsler, "Evolutionary pursuit and its application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 570C582, 2000
- [82] C. Liu and H. Wechsler, "Robust coding scheme for indexing and retrieval from large face databases," *IEEE Transactions on Image Process*, vol. 9, pp. 132C137, 2000
- [83] M.D. Kelly, "Visual identification of people by computer," Tech. rep. AI-130, Stanford AI Project, Stanford, CA
- [84] T. Kanade, "Computer recognition of human faces," Birkhauser, Basel, Switzerland, and Stuttgart, Germany, 1973
- [85] I.J. Cox, J. Ghosn and P. N. Yianilos, "Feature-based face recognition using mixture distance," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 209C216, 1996
- [86] A.V. Nefian and M.H. Hayes III, "Hidden Markov models for face recognition," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 2721C2724, 1998
- [87] F. Samaria, "Face recognition using hidden markov models," Ph.D. dissertation, University of Cambridge, Cambridge, U.K., 1994
- [88] F. Samaria and S. Young, "HMM-based architecture for face identification," *Image and Vision Computing*, vol. 12, pp. 537C583, 1994
- [89] K. Okdada, J. Steffans, T. Maurer, H. Hong, E. Elagin, H. Neven, C.V.D. Andmalsburg, "The Bochum/USC Face Recognition System and how it fared

- in the FERET Phase III Test," *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, pp. 186C205, 1998
- [90] L. Wiskott, J.-M. Fellous^{*} and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775C779, 1997
- [91] J. Buhmann, M. Lades and C. von der Malsburg, "Size and distortion invariant object recognition by hierarchical graph matching," in *Proceedings of International Joint Conference on Neural Networks*, pp. 411C416, 1990
- [92] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. v. d. Malsburg, R. P. Wurtz and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Transactions on Computers*, vol. 42, pp. 300C311, 1993
- [93] N. Kruger, M. Potzsch and C. v. d. Andmalsburg, "Determination of face position and pose with a learned representation based on labelled graphs," *Image and Vision Computing*, vol. 15, 665C673, 1997
- [94] T. Maurer and C. v. d. Andmalsburg, "Single view based recognition of faces rotated in depth," in *Proceedings of International Workshop on Automatic Face and Gesture Recognition*, pp. 176C181, 1996
- [95] S. Lawrence, C. L. Giles, A. C. Tsoi and A. D. Back, "Face Recognition: A Convolutional Neural Network Approach," *IEEE Transactions on Neural Networks*, vol. 8, pp. 98C113, 1997
- [96] P. Penev and J. Atick, "Local feature analysis: A general statistical theory for object representation," *Network: Computation in Neural Systems* vol. 7, pp. 477C500, 1996

- [97] A. Lanitis, C.J. Taylor and T.F. Cootes, "Automatic face identification system using flexible appearance models," *Image and Vision Computing*, vol. 13, pp. 393C401, 1995
- [98] J. Huang, B. Heisele and V. Blanz, "Component-based face recognition with 3D morphable models," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 5, pp. 85-92, 2004
- [99] B. Heisele, T. Serre, M. Pontil and T. Poggio, "Component-based face detection," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, 2001
- [100] V. Blanz and T. Vetter, "A Morphable model for the synthesis of 3D faces," in Proceedings of *SIGGRAPH99*, pp. 187C194, 1999
- [101] P. Penev and L. Sirovich, "The global dimensionality of face space," in Proceedings of *International Conference on Automatic Face and Gesture Recognition*, pp. 264-270, 2000
- [102] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA, 1995
- [103] M.S. Bartlett, H.M. Lades and T. Sejnowski, "Independent component representation for face recognition," in Proceedings of *SPIE Symposium on Electronic Imaging: Science and Technology*, pp. 528C539, 1998
- [104] A.J. Bell and T.J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp.1129C1159, 1995
- [105] A.J. Bell and T.J. Sejnowski, "The independent components of natural scenes are edge filters," *Vision Research*, vol. 37, pp. 3327C3338, 1997
- [106] I. Biederman and P. Kalocsai, "Neural and psychophysical analysis of object and face recognition," *Face Recognition: From Theory to Applications*,

- H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, pp. 3C25, 1998
- [107] V. Bruce, P.J.B. Hancock and A.M. Burton, "Human face perception and identification," *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, 51C72, 1998
- [108] B. Knight and A. Johnston, "The role of movement in face recognition," *Visual Cognition*, vol. 4, pp. 265-274, 1997
- [109] T. Choudhury, B. Clarkson, T. Jebara and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in Proceedings of *International Conference on Audio- and Video-Based Person Authentication*, pp. 176C181, 1999
- [110] S. McKenna and S. Gong, "Recognizing moving faces," *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, Eds. Springer-Verlag, Berlin, Germany, pp. 578C588, 1998
- [111] S. Gong, S. McKenna and A. Psarrou, *Dynamic Vision: From Images to Face Recognition*, World Scientific, Singapore, May 2000
- [112] L. Gu, S.Z. Li and H.J. Zhang, "Learning probabilistic distribution model for multiview face detection," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 116-122, 2001
- [113] Y. Li, S. Gong and H. Liddell, "Modelling face dynamics across view and over time," in Proceedings of *International Conference on Computer Vision*, 2001
- [114] B. Li and R. Chellappa, "Face verification through tracking facial features," *Journal of Optical Society of America*, vol. 18, 2001
- [115] Y. Li, S. Gong and H. Liddell, "Constructing facial identity surfaces in a non-linear discriminating space," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 258-264, 2001

- [116] D. Decarlo and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *International Journal of Computer Vision*, vol. 38, pp. 99C127, 2000
- [117] H. Wechsler, V. Kakkad, J. Huang, S. Gutta and V. Chen, "Automatic video-based person authentication using the RBF network," in Proceedings of *International Conference on Audioand Video-Based Person Authentication*, pp. 85C92, 1997
- [118] S. McKenna and S. Gong, "Non-intrusive person authentication for access control by visual tracking and face recognition," in Proceedings of *International Conference on Audio- and Video-Based Person Authentication*, pp. 177C183, 1997
- [119] G.J. Edwards, C.J. Taylor and T.F. Cootes, "Learning to identify and track faces in image sequences," in Proceedings of *International Conference on Automatic Face and Gesture Recognition*, 1998
- [120] T. Cootes, C. Taylor, D. Cooper and J. Graham, "Active shape models – their training and application," *Computer Vision and Image Understanding*, vol. 61, pp. 18C23, 1995
- [121] S. Zhou, V. Krueger and R. Chellappa, "Probabilistic recognition of human faces from video," *Computer Vision and Image Understanding*, vol. 91, pp. 214C245, 2003
- [122] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K.-R. Muller, "Fisher discriminant analysis with kernels," in Proceedings of *IEEE Workshop on Neural Networks for Signal Processing*, 1999
- [123] B.K.P. Horn and M.J. Brooks, *Shape from Shading*. MIT Press, Cambridge, MA, 1989

- [124] Y. Adini, Y. Moses and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 721C732, 1997
- [125] K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 39C51, 1997
- [126] D.W. Jacobs, P.N. Belhumeur and R. Basri, "Comparing images under variable illumination," in proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 610C617, 1998
- [127] P.N. Belhumeur and D.J. Kriegman, "What is the set of images of an object under all possible lighting conditions?" in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 52C58, 1997
- [128] A. Shashua and T. R. Raviv. "The quotient image: Class based re-rendering and recognition with varying illuminations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 129–139, 2001
- [129] W.T. Freeman and J.B. Tenenbaum, "Separating style and contents with bilinear models," *Neural Computation*, vol. 12, pp. 1247C1283, 2000
- [130] J. Atick, P. Griffin and N. Redlich, "Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images," *Neural Computation*, vol. 8, pp. 1321C1340, 1996
- [131] P.S. Tsai and M. Shah, "Shape from shading using linear approximation," *Image and Vision Computing*, vol. 12, pp. 487C498, 1994
- [132] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," in proceedings of *International Conference on Computer Vision*, vol. 2, pp. 383C390, 2001

- [133] D.J. Beymer, "Face recognition under varying pose," Tech. Rep. 1461. MIT AI Lab, Massachusetts Institute of Technology, Cambridge, MA, 1993
- [134] S. Ullman and R. Basri, "Recognition by linear combinations of models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 992C1006, 1991
- [135] D.J. Beymer and T. Poggio, "Face recognition from one example view," in Proceedings of *International Conference on Computer Vision*, pp. 500C507, 1995
- [136] T. Cootes, K. Walker and C. Taylor, "View-based active appearance models," in Proceedings of *International Conference on Automatic Face and Gesture Recognition*, pp. 227-238, 2000
- [137] E. Sali and S. Ullman, "Recognizing novel 3-D objects under new illumination and viewing position using a small number of example views or even a single view," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 153C161, 1998
- [138] T. Vetter and T. Poggio, "Linear object classes and image synthesis from a single example image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 733C742, 1997
- [139] H. Kong, E.K. Teoh, J.G. Wang and V. Ronda, "Two Dimensional Fisher Discriminant Analysis: Forget about the Small Sample Size Problem," in Proceedings of *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.2, pp.761-764, USA, 19-23 Mar 2005
- [140] H. Kong and L. Wang and E.K. Teoh and J.G. Wang and R. Venkateswarlu, "A Framework of 2D Fisher Discriminant Analysis: Application to Face Recognition with Small Number of Training Samples," in Proceedings of *IEEE Conference on Computer Vision and Pattern Recognition*, vol.2, pp. 1083-1088, 2005

- [141] L. Breima, "Bagging predictors", *Machine Learning*, pp.123-140, 1996
- [142] R.E. Schapire and Y. Singer, "Improved Boosting Algorithms Using Confidence-rated Predictions", *Machine Learning*, pp. 297-336, 1999
- [143] L.I. Kuncheva and C.J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181-207, 2003
- [144] R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955-966, 1995
- [145] Y.S. Huang and C.Y. Suen, "A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 1, pp. 90-94, 1995
- [146] W.P. Kegelmeyer and K. Bowyer, "Combination of Multiple Classifier Using Local Accuracy Estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 405-410, 1997
- [147] J. Kitter, M. Hatef, R.P.W. Duin and J. Matas, "On Combining Classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998
- [148] S.B. Yacoub, Y. Abdeljaoud and E. Mayoraz, "Fusion of Face and Speech Data for Person Identity Verification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1065-1074, 1999
- [149] M. Skurichina, L.I. Kuncheva and R.P.W. Duin, "Bagging, Boosting, and the Random Subspace Method for Linear Classifier," *Pattern Analysis and Applications*, vol. 5, no. 2, pp. 121-135, 2002
- [150] T. Windeatt, "Diversity measures for multiple classifier system analysis and design," *Information Fusion*, vol. 6, pp. 21-36, 2005

- [151] T.G. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization", *Machine Learning*, vol. 40, No. 2, pp. 139-157, 2000.
- [152] D.W. Opitz, "Feature Selection for Ensembles", in Proceedings of the *Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence*, pp.379-384, 1999
- [153] G. Brown and J.L. Wyatt, "The Use of the Ambiguity Decomposition in Neural Network Ensemble Learning Methods," in Proceedings of *International Conference on Machine Learning*, pp.67-74, 2003
- [154] H. W. Shin and S. Y. Sohn, "Selected tree classifier combination based on both accuracy and error diversity," *Pattern Recognition*, vol.38, pp.191-197, 2005
- [155] ORL face database, [http : //www.uk.research.att.com/ facedatabase.html](http://www.uk.research.att.com/facedatabase.html), AT&T Laboratories, Cambridge, UK
- [156] CMU PIE face database, [http : //www.ri.cmu.edu/projects/project418.html](http://www.ri.cmu.edu/projects/project418.html), Carnegie Mellon University Robotics Institute, Pittsburgh, U.S.
- [157] The Yale Face Database, [http : //cvc.yale.edu/projects/yalefaces/yalefaces.html](http://cvc.yale.edu/projects/yalefaces/yalefaces.html), Yale University, New Haven, U.S.
- [158] The Yale Face Database B, [http : //cvc.yale.edu/projects/yalefacesB/yalefacesB.html](http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html), Yale University, New Haven, U.S.
- [159] The UMIST face database, [http : //images.ee.umist.ac.uk/danny/database.html](http://images.ee.umist.ac.uk/danny/database.html), UMIST, UK
- [160] The CMU AMP Expression face database, [http : //amp.ece.cmu.edu/downloads.htm](http://amp.ece.cmu.edu/downloads.htm), Advanced Multimedia Processing Lab , Carnegie Mellon University, Pittsburgh, U.S.