Enhanced linear subspace methods for face and gait analysis

Lu, Jiwen

2011

Lu, J. (2011). Enhanced linear subspace methods for face and gait analysis. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/46438>

<https://doi.org/10.32657/10356/46438>

Enhanced Linear Subspace Methods for Face and Gait Analysis

Lu Jiwen

School of Electrical & Electronic Engineering

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

2011

Acknowledgments

I would like to express my deep and sincere gratitude to my supervisor, Prof. Yap-Peng Tan. His enthusiastic supervision and invaluable guidance have been essential for the results presented here. I am very grateful that he has spent so much time with me to discuss different research problems. His knowledge, suggestions, and discussions help me to become a more capable researcher. His encouragement also helps me to overcome the difficulties encountered in my research.

I would also like to thank my seniors and friends, Dr Gao Yang and Dr Ji Tao, who have been very supportive in discussing with me on my research problems.

Lastly, my sincere thanks to the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, for providing me the full research scholarship and support to complete my Ph.D study.

Contents

Acknowledgments	i
Contents	ii
Summary	vi
List of Figures	viii
List of Tables	xi
1 Introduction	1
1.1 Motivations	1
1.2 Thesis Contributions	5
1.3 Thesis Organization	7
2 Literature Review	8
2.1 General Overview on Subspace Learning	8
2.2 Literature Review	9
2.2.1 PCA	9
2.2.2 ICA	10
2.2.3 LDA	11
2.2.4 LPP	12
2.2.5 DLPP	13
2.2.6 MFA	14
2.2.7 NMF	15

3	Doubly Weighted Subspace Learning	17
3.1	Introduction	17
3.2	Doubly Weighted Subspace Learning	18
3.2.1	Motivation	18
3.2.2	Between-Sample Weighting Matrix	20
3.2.3	Within-Sample Weighting Matrix	22
3.2.4	Proposed Algorithms	25
3.2.5	Relation with Existing Weighted Techniques	26
3.3	Experiments and Results	27
3.3.1	Data Preparation	27
3.3.2	Face Representation	28
3.3.3	Face Recognition	31
3.3.4	Discussion	34
3.4	Chapter Summary	35
4	Parametric Regularized Subspace Learning	36
4.1	Introduction	36
4.2	Parametric Regularized Subspace Learning	38
4.3	Experimental Results	44
4.3.1	Locality Preserving Power	45
4.3.2	Face Recognition	45
4.3.3	Discussion	50
4.4	Chapter Summary	52
5	Cost-Sensitive Subspace Learning	53
5.1	Introduction	53
5.2	Cost-Sensitive Subspace Learning	55
5.2.1	Motivation	55
5.2.2	Proposed Methods	56
5.2.3	Discussion	59
5.3	Experimental Results	60
5.3.1	Toy Example	61

5.3.2	Face Representation and Recognition	62
5.3.3	Discussion	69
5.4	Chapter Summary	70
6	Subspace Learning with Limited Number of Training Samples	71
6.1	Introduction	71
6.2	NFL and NFS	73
6.2.1	NFL	73
6.2.2	NFS	73
6.3	Proposed Methods	74
6.3.1	DNFLA	74
6.3.2	NFSA	76
6.3.3	DNFSA	77
6.4	Experimental Results	80
6.4.1	Face Recognition	80
6.4.2	Facial Expression Recognition	81
6.4.3	Computational Complexity	84
6.4.4	Discussion	85
6.5	Chapter Summary	85
7	Multi-Label Subspace Learning for Human Age Estimation	88
7.1	Introduction	88
7.2	Proposed Approach	91
7.2.1	Feature Extraction	91
7.2.2	Label Encoding	92
7.2.3	MLG Subspace Learning	92
7.2.4	Recognition and Age Decoding	93
7.3	Multiple Feature Fusion	94
7.4	Experiments and Results	97
7.4.1	Data Preparation	97
7.4.2	Results	98
7.4.3	Discussion	105
7.5	Chapter Summary	106

8	Conclusions and Future Work	107
8.1	Conclusions	107
8.2	Future Work	109
	Author's Publications	111
	Bibliography	114

Summary

Feature extraction has been extensively investigated and discussed in computer vision and pattern recognition literature over the past three decades. It has particularly attracted more and more attention in recent years due to the increasing demands for developing real-world human computer interaction systems. While a large number of feature extraction algorithms have been proposed in the literature and some of them have achieved reasonably good performance in many real world applications, such as face recognition, gait recognition, facial expression recognition and human age estimation, there is still some room for further improvement to address the challenges of these methods. In this thesis, we propose various enhanced linear subspace algorithms and apply them to face and gait feature extraction to demonstrate their efficacy and superiority over state-of-the-art methods.

Motivated by the fact that some face samples and parts are more effectual in feature extraction, we construct two weighting matrices based on pairwise similarity of face samples within a same class and discriminant score of each pixel within a face sample to duly emphasize both the between-sample and within-sample face image features. We then incorporate these two weighting matrices into three popular subspace learning methods to obtain the discriminative features for face recognition. Since most existing subspace learning algorithms usually suffer from the small sample size (SSS) problem in many feature extraction applications, especially for face recognition which stems from the fact that the number of training images is usually much smaller than the dimension of each face image, we propose a parametric regularization approach to perform feature extraction in a whole feature space rather than a reduced principal component analysis (PCA) subspace for face recognition. Through this, a more locality preserving power and better recognition accuracy are achieved. Inspired by the fact that different misclassifications

usually lead to different amounts of loss for a practical classification system, we propose a cost-sensitive subspace learning approach for feature extraction. Our approach incorporates a cost matrix, which specifies the different costs associated with misclassifications of different subjects, into three popular subspace learning algorithms to obtain the corresponding cost-sensitive methods, achieving a minimum overall recognition loss by performing recognition in the low-dimensional subspaces derived. Recent advances have shown that the performance of subspace learning algorithms degrade heavily when the number of training samples per class is small because such limited samples are usually not able to effectively characterize the geometrical structure of the original sample space for learning an effective feature subspace. To address this problem, we propose using the nearest feature line (NFL) and nearest feature space (NFS) metrics rather than the conventional nearest neighbor (NN) metric to seek discriminative feature spaces for feature extraction. To better correlate multiple label information of samples, we propose learning a multi-label-guided (MLG) subspace to better correlate these multi-label information. To enhance the robustness of the proposed method, we extract a set of over-complete Gabor features including both Gabor magnitude (GM) and Gabor phase (GP) information of the samples and perform multiple feature fusion to further improve the performance.

The above-mentioned methods have been successfully applied to several computer vision applications, such as face recognition, facial expression recognition and gait-based human age estimation. Experimental results are presented to demonstrate the efficacy of the proposed methods.

List of Figures

3.1	An example showing the similarities and differences among samples in each class.	21
3.2	Partitioning a face image into patches.	23
3.3	Visualization of a within-sample weighting matrix.	24
3.4	Flowcharts of proposed doubly weighted approach.	25
3.5	Aligned and resized face samples of one subject in the (a) Extended Yale B database and (b) PIE database.	28
3.6	Set of the first 10 basis images of (a) PCA, (b) LDA, (c) NMF, (d) DW-PCA, (e) DWLDA, and (f) DWNMF methods.	29
3.7	Normalized kurtosis of NMF and DWNMF.	30
3.8	Performance of proposed DWPCA, DWLDA and DWNMF methods versus different region sizes.	33
4.1	Spectrum of F	40
4.2	The real spectrum and the modeled spectrum with different value of regularization parameter a	41
4.3	Recognition accuracy versus different value of parameter b	42
4.4	Sample face images from the ORL database.	44
4.5	Sample face images from the FERET database.	45
4.6	Eigenvalues of LPP and PRLPP. The eigenvalues of PRLPP are consistently smaller than those of LPP, which indicates that PRLPP has more locality preserving power than LPP.	46
4.7	Performance comparison on three different spaces.	47
4.8	ARA comparison of PRLPP with PCA, LDA, and LPP.	48

4.9	ARA versus different number of training samples of each subject.	48
4.10	Performance comparison with different number of training samples per subject.	50
5.1	The projection direction sought by LDA and CSLDA, respectively.	62
5.2	Samples of the cropped and aligned face image from the (a) AR database and (b) FERET database.	63
5.3	First 10 basis images of (a) PCA, (b) LDA, (c) LPP, (d) CSPCA, (e) CSLDA, and (f) CSLPP methods.	64
5.4	Comparison on total cost with different number of gallery subjects. The first row are the results on the AR database. The second row are the results on the FERET database.	66
5.5	Total cost on the AR database of different cost-sensitive subspace learning methods when the uniform-cost and cost-sensitive classifiers are applied, respectively.	67
5.6	Comparison on total cost with different number of gallery subjects. The first row are the results on the AR database. The second row are the results on the FERET database.	69
6.1	The feature line $\overline{x_1x_2}$ of two samples, a probe sample x_p , and its projection on the feature line, x_q	74
6.2	The feature space distance between a probe sample x_p , and its projection on the feature space, x_q	75
6.3	A simple example illustrating the subspace projection by the NFSA and DNFSFA methods: (a) Original data, (b) NFSA, (c) DNFSFA.	78
6.4	ARAs achieved by various methods under different feature dimensions. The results obtained when (a) the NN classifier, (b) the NFL classifier, and (c) the NFS classifier are applied, respectively.	82
6.5	Samples of face images of a subject from the C-K database with different expressions. (a) Surprise. (b) Fear. (c) Sadness. (d) Anger, (e) Disgust, and (f) Happy.	83

6.6	ARAs achieved by various methods under different feature dimensions. The results obtained when (a) the NN classifier, (b) the NFL classifier, and (c) the NFS classifier are applied, respectively.	87
7.1	Gait sequences of two persons with the same age (26 years old) but different gender. (a) Male and (b) Female.	90
7.2	Flow chart of the proposed gait-based age estimation.	91
7.3	Normalized and aligned silhouette images of a gait sequence. The right-most image is the GEI.	92
7.4	(a) Real parts of 40 Gabor kernels. (b) GM representation of a GEI. (c) GP representation of a GEI.	95
7.5	Flow chart of multiple feature fusion-based age estimation framework. . .	96
7.6	Aging manifold learned by the (a) PCA method on the Gallery set, (b) CEA method on the Gallery set, (c) PCA method on the Probe A set and (d) CEA method on the Probe A set, respectively. The color data points of age from 19 to 59 are colored from blue to red.	101
7.7	Age range distribution of the selected USF subset database.	102
7.8	MAEs of estimation for three different methods when different percentages of training samples have missing gender labels on the gallery and probe A subsets of the USF database.	103
7.9	Sample frames of one subject from our gait database.	105

List of Tables

2.1	Categorization of subspace methods according to different data structures.	9
3.1	Average pairwise mutual information (APMI) of six subspace learning methods.	29
3.2	Number of samples in the four training sets from the Extended Yale B and PIE Databases.	31
3.3	Recognition performance comparison on the Extended Yale B database (mean $\pm$ std).	32
3.4	Recognition performance comparison on the PIE database (mean $\pm$ std). .	32
3.5	Comparison with existing weighted methods on the Extended Yale B database. (mean $\pm$ std).	34
4.1	S_1 and S_2 for LPP and its three discriminant extensions.	37
4.2	Comparison on recognition accuracy using the ORL and FERET databases.	47
4.3	Comparison with existing regularization methods.	49
4.4	Performance comparison of three manifold learning algorithms and our proposed regularized extensions.	49
4.5	CPU time (in second) used by different methods on the ORL database. .	49
5.1	Cost matrix of a practical face recognition system.	56
5.2	Cost matrix used in the toy example. Class 1, 2, and 3 denote the cross, circle and square class, respectively.	61
5.3	Comparison of total cost and total error rate (TER) on synthetic data set.	62
5.4	Experimental settings for face representation.	64
5.5	Parameter settings for a face recognition-based access control system. . .	65

5.6	Comparison of total cost (cost), total error rate (TER), false rejection rate (FRR) and false acceptance rate (FAR) on the AR and FERET databases for the access control application.	65
5.7	CPU time (in second) used by different methods on the AR database. . .	67
5.8	Parameter settings for a face recognition-based criminal search system. .	68
5.9	Comparison of total cost (cost), total error rate (TER), false rejection rate (FRR) and false acceptance rate (FAR) on the AR and FERET databases for the criminal search application.	68
6.1	Comparison on ARA (%) of different algorithms using the AR database.	81
6.2	Comparison on ARA (%) of different algorithms using the C-K database.	84
6.3	CPU time (in second) used by different methods on the AR database. . .	84
7.1	Detailed information of all the subsets of the USF database.	97
7.2	MAEs (years old) on the USF gait database using different features. . . .	99
7.3	MAEs (years old) on all the 13 sets of the USF database using different methods.	100
7.4	MAEs (years old) on the two subsets of the USF database using different methods.	103
7.5	MAEs (years old) on the two subsets of the USF database using different gait features.	104
7.6	MAEs (years old) on our dataset using different gait features.	105

Chapter 1

Introduction

1.1 Motivations

In many areas of information processing such as machine learning, information retrieval, data mining, computer vision and pattern recognition, one is often confronted with high-dimensional data. More often than not, it is necessary to transform these high-dimensional data to a low-dimensional feature space because high-dimensional data usually result in high computational complexity and deteriorate the classification performance of classifiers. Feature extraction with subspace learning is one of the most effective techniques to address such problems, and it has been extensively investigated and discussed in many areas such as biometrics, image and video retrieval, bioinformatics, and visual surveillance over the past three decades. More recently, subspace learning has attracted more and more attention due to the increasing demands of developing real-world human computer interaction and intelligent visual surveillance systems. The common aim of different subspace learning methods is to learn one or multiple compact, low-dimensional feature subspace(s) for data representation, in such a way that some intrinsic characteristics of the original data are well preserved.

Two representative and popular subspace learning algorithms are principal component analysis (PCA) [1, 3] and linear discriminant analysis (LDA) [2, 3]. PCA performs dimensionality reduction and feature extraction by projecting the original high-dimensional data to a low-dimensional linear subspace spanned by the leading eigenvectors of the covariance matrix, leading to a set of mutually orthogonal bases that retain the maximum variance of the samples. LDA is a supervised subspace learning approach which searches

for a set of most discriminant projections to maximize the ratio of between-class variance to within-class variance. As label information has been exploited in the learning phase, LDA usually outperforms PCA for classification tasks. While these two algorithms have attained reasonable performance in many tasks, such as face and gait recognition [1, 2, 131–134], they may fall short in uncovering the underlying nonlinear manifold of high-dimensional these samples as they seek only a compact Euclidean subspace for efficient data representation and recognition.

Over the past recent years, we have witnessed a considerable interest in manifold-based subspace learning algorithms for feature extraction. The basic idea of these methods is that high-dimensional samples can be considered as a set of geometrically related points lying on or nearby a smooth, low-dimensional manifold [5–27]. The most representative such algorithm is locality preserving projections (LPP) [8]. LPP aims to seek a low-dimensional subspace that preserves the intrinsic geometric structure of the original samples, allowing samples lying in a neighborhood to maintain their locality relationship after projection. As LPP is originally unsupervised, more recent work has included supervised information into the formulation and derived many discriminant manifold learning algorithms, such as discriminant locality preserving projections (DLPP) [20], margin fisher analysis (MFA) [17], local discriminant embedding (LDE) [27], discriminant simplex analysis (DSA) [18, 160], and neighborhood preserving discriminant projections (NPDP) [26], to enhance the recognition performance. Despite the different assumptions of these methods, they can be unified into a general graph embedding (GE) framework [13, 14, 17] with different constraints.

While the above subspace learning methods have achieved reasonable success in many feature extraction tasks, there are still several issues to be addressed:

- 1) Most existing subspace learning algorithms consider each sample holistically and ignore the fact that different samples may have different representative power for feature extraction and recognition. For example, most existing subspace learning methods consider each sample in the training set contributing equally in the training phase, i.e., the contribution of each sample is usually set to be $1/N$, where N is the number of the training samples. However, this assumption of equal contribution does not necessarily hold in real-world applications. For example, when face images are sampled from a video

sequence during the training phase, it is possible to have many similar face samples which may not necessarily reflect the distribution of the faces acquired in the testing (recognition) phase. Hence, it may be more reasonable to assign different weight to each face sample to reflect its different importance during the training. Moreover, in many practical recognition applications, it is difficult and expensive to collect enough training samples, and we need to exploit the discriminant information from limited number of training samples. On the other hand, it is likely that different pixels in different spatial positions of each face sample may have different discriminative power. Most existing subspace learning methods extract features based on some statistical analysis and do not consider the physical meaning of the features such as the eyes and mouths of the face samples. That is, each element in a face image is treated equally and acts the same in subspace learning. However, due to the different spatial importance of each pixel in a face image, the pixel with different values and spatial locations could be assigned different weights for discrimination. For example, pixels at important facial features such as eyes, mouth, nose, etc., could be emphasized and others such as cheek and forehead could be deemphasized. Is it possible to give different emphases on different samples according to their different importance in subspace learning?

2) Conventional subspace learning algorithms usually suffer from the small sample size (SSS) problem in many image classification tasks such as face recognition, and most of them mainly apply PCA to mitigate this problem. However, some important discriminant information could be lost in the PCA preprocessing stage. Is there an effective subspace learning approach which can extract discriminative features for recognition from a whole feature space rather than a reduced subspace of PCA?

3) Conventional subspace learning methods aim to seek low recognition errors and implicitly assumes same loss for all misclassifications. In many real-world applications, however, this assumption may not hold as different misclassifications could lead to different losses. Existing subspace learning-based face recognition algorithms aim to seek low recognition errors and implicitly assume same loss of all misclassifications. In many real-world face recognition applications, however, this assumption may not hold as different misclassifications could lead to different losses. For example, it may cause inconvenience to a gallery person who is mis-recognized as an impostor and not allowed to enter the

room by a face recognition-based door-locker, but may result in a serious loss or damage if an impostor is mis-recognized as a gallery person and allowed to enter the room. Another example is the face sketch recognition-based criminal search system. It could cause some trouble to an innocent citizen who is mis-recognized as a criminal suspect by the criminal search system, but may result in a large insecurity if a criminal suspect is mis-recognized as innocent citizen and allowed to go free. From these two typical application examples, we can conclude that face recognition is a cost-sensitive pattern classification problem and it is not very reasonable, at least not optimal, to only seek the minimal recognition error in many practical face recognition applications. Is it possible to learn a cost-sensitive subspace which explicitly includes cost information of each sample for recognition?

4) Most existing subspace learning algorithms apply the conventional nearest neighbor (NN) metric to derive the subspace and may not effectively characterize the geometrical information of samples, especially when the number of training samples per class is limited. In many practical pattern classification applications, such as face recognition, it is difficult and expensive to collect a large number of training samples, and we need to fully exploit the discriminant information from the available limited training samples. Aiming at generalizing the representative capacity of the training samples, Li and Lu [70] proposed a novel nearest feature line (NFL) method to virtually enlarge the training set for face recognition. The basic idea of NFL is using a linear model to interpolate and extrapolate each pair of prototypes belonging to the same class to model the possible variants of the training samples. Because of this excellent generalization capacity, NFL has been widely used in many computer vision and pattern recognition applications, such as face recognition [70, 74, 75], audio classification [71, 72, 76], and image retrieval [72]. Is it possible to learn an effective subspace in case of limited training per class?

5) Most conventional subspace learning algorithms utilize single label information of the samples for feature extraction. In some applications, there could be some other additional useful label information, and it is likely that some data belongs to multiple classes, causing the actual classes to be overlapped. For example, in medical diagnosis, a disease may belong to multiple categories, and genes may have multiple functions, yielding multiple labels. That is, in multi-label classification, class memberships can

be inferred from one another through label correlations, which provide an important opportunity to improve classification accuracy. As a result, a multi-label classification method should make use of label correlations for improved classification performance. Is it possible to make use of both multi-label information to improve the performance? If so, how to achieve this goal?

1.2 Thesis Contributions

Aiming at addressing the aforementioned problems, we propose in this thesis several enhanced linear subspace methods, namely doubly weighted subspace learning, parametric regularized subspace learning, cost-sensitive subspace learning, subspace learning with limited number of training samples, and multi-label subspace learning. The main contributions of this thesis are summarized below.

1) We propose a doubly weighted subspace learning approach for face representation and recognition. Motivated by the fact that some face samples and parts are more effectual in characterizing and recognizing faces, we construct two weighting matrices based on pairwise similarity of face samples within a same class and discriminant score of each pixel within a face sample to duly emphasize both the between-sample and within-sample features. We then incorporate these two weighting matrices into three popular subspace learning methods, namely PCA, LDA and nonnegative matrix factorization (NMF), to obtain the discriminative features of faces for recognition. Experimental results show that the proposed approach can effectively enhance the discriminant power of the extracted face features and outperform existing, non-weighted subspace learning algorithms. The performance gain is even more apparent for cases with imbalanced number of training samples. (Chapter 3)

2) We propose a parametric regularized subspace learning approach for face recognition. Our objective is to regulate the LPP feature space in a parametric manner, and extract useful discriminant information from the whole feature space rather than a reduced PCA subspace. This results in better locality preserving power and higher recognition accuracy than the original LPP method. Moreover, the proposed regularization method can be easily extended to other newly proposed subspace learning algorithms

and to effectively address the SSS problem. Experimental results on two widely used face databases demonstrate the efficacy of the proposed regularization method and the superiority over existing solutions to the SSS problem. (Chapter 4)

3) We propose a cost-sensitive subspace learning approach for face representation and recognition. Our approach incorporates a cost matrix, which specifies the different costs associated with misclassifications of different subjects, into three popular subspace learning algorithms and devise the corresponding cost-sensitive subspace learning methods, namely, cost-sensitive principal component analysis (CSPCA), cost-sensitive linear discriminant analysis (CSLDA), and cost-sensitive locality preserving projections (CSLPP), to achieve a minimum overall recognition loss by performing recognition in the low-dimensional subspace derived. Experimental results are presented to demonstrate that our proposed approach outperforms existing uniform-cost subspace learning methods. (Chapter 5)

4) We propose using the nearest feature line (NFL) and nearest feature space (NFS) metrics rather than the nearest neighbor (NN) metric to seek discriminative feature spaces such that within-class feature line (FL) and feature space (FS) distances are minimized and between-class FL and FS distances are maximized simultaneously in the reduced subspace. Based on this criterion, we propose three new subspace learning algorithms, called discriminant nearest feature line analysis (DNFLA), nearest feature space analysis (NFSA), and discriminant nearest feature space analysis (DNFSA). Experimental results are presented to demonstrate that these proposed methods consistently outperform existing conventional NN-based subspace methods. (Chapter 6)

5) We apply the multi-label subspace learning technique to address the problem of gait-based human age estimation. Motivated by the fact that human gait appearances vary between males and females even within the same age group, we learn a multi-label-guided (MLG) subspace to better characterize and correlate the age and gender information of a person for estimating human age. As human age assumes only nonnegative value and existing multi-label learning techniques mainly deal with ensembles of binary classes, we devise an effective label encoding scheme to convert each age value to a binary sequence, making conventional multi-label learning suitable for our task. To better characterize human gait appearance and enhance the robustness of the proposed

age estimation method, we extract a set of over-complete Gabor features including both Gabor magnitude (GM) and Gabor phase (GP) information of a gait sequence and perform multiple feature fusion to enhance the age estimation performance. Experimental results clearly demonstrate the feasibility of using gait signatures to estimate human age and the efficacy of our proposed method. (Chapter 7)

1.3 Thesis Organization

This thesis is organized as follows: Chapter 1 presents the motivations and the major contributions of this thesis. Chapter 2 provides a general overview of existing feature extraction (subspace learning) techniques. In Chapters 3-5, we present the double weighted subspace learning, regularized subspace learning and cost-sensitive subspace learning methods, respectively, for face recognition. Chapter 6 proposes one NFL-based and two NFS-based discriminative subspace learning algorithms. Chapter 7 applies the multi-label subspace learning technique to address the problem of gait-based human age estimation. Finally, conclusions and future work are presented in Chapter 8.

Chapter 2

Literature Review

While many algorithms have been proposed for face and gait analysis, the most successful and popular approaches to date are generally the appearance-based methods with subspace learning [1–4]. In this chapter, we briefly review some representative subspace learning-based face and gait analysis algorithms.

2.1 General Overview on Subspace Learning

Subspace learning is the process of mapping high-dimensional data to a lower dimensional subspace, which is typically used as a preprocessing step in many pattern classification applications. Subspace learning is one of the most effective techniques for feature extraction, and a large number of such algorithms have been proposed over the past two decades [1–34, 36–48]. The common objective of subspace learning is to seek, in a supervised, unsupervised or semi-supervised manner, one or multiple succinct low-dimensional feature space(s) to preserve some intrinsic characteristics of the original data samples.

Several criteria can be used to classify subspace learning algorithms into different categories. According to the label information of the samples used in the training phase, existing subspace learning approaches can be classified into unsupervised, semi-supervised and supervised methods, corresponding to utilizing no label, partial label and full label information of the training samples, respectively. According to the different data structures of the samples, existing subspace learning methods can also be classified into linear subspace, bilinear subspace and multilinear subspace methods, respectively. Moreover, they can also be categorized into linear and nonlinear methods depends on

Table 2.1: Categorization of subspace methods according to different data structures.

Method	Representative algorithms
Linear	PCA [1], ICA [151, 159], LDA [2], LPP [8], DLPP [20], MFA [17], NMF [36, 37]
Bilinear	BDPCA [34], 2DLDA [29], 2DLPP [30], (2D) ² DLPP [31]
Multilinear	MPCA [32], MDA [33], MLPP [17], TMFA [17]
Nonlinear	KPCA [4], KLDA [4], KLPP [17], KMFA [17]

whether the mapping function is linear or nonlinear. Table 2.1 lists several representative subspace learning algorithms of linear, bilinear, multilinear and nonlinear methods.

In this thesis, we focus on linear subspace methods due to the following reasons:

1) Simplicity: linear subspace methods are generally much easier to be implemented than other nonlinear methods because neither a kernel function nor a tensor model is pre-specified;

2) Closed-form solution: the mapping function of linear subspace methods can usually be analytically obtained and the bases of the mapping function can be easily visualized;

3) Efficacy: the performance of linear subspace methods are comparable to many other nonlinear, bilinear and multilinear methods.

Hence, we focus on linear subspace methods in this thesis and review several representative linear subspace learning methods in the following.

2.2 Literature Review

Consider a set of samples denoted $X = [x_1, x_2, \dots, x_N]$, $\{x_i \in R^d, i = 1, 2, \dots, N\}$, and the corresponding label $\{l_i \in \{1, 2, \dots, c\}, i = 1, 2, \dots, N\}$, where d and l_i is the feature dimension and label of x_i , respectively. The objective of linear subspace learning is to find a linear mapping $W = [w_1, w_2, \dots, w_k]$ to project $\{x_i, i = 1, 2, \dots, N\}$ into a low dimensional representation $\{y_i \in R^m, i = 1, 2, \dots, N\}$, i.e., $y_i = W^T x_i$, $m \ll d$. The essential difference of these linear subspace methods lies in the difference in defining and finding the mapping W .

2.2.1 PCA

PCA seeks to find a set of projection axes such that the global scatter is maximized after the projection of the samples. The global scatter can be characterized by the mean

square of the Euclidean distance between any pair of the projected sample points, defined as [1]

$$\max J_T(w) = \frac{1}{2} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \|y_i - y_j\|^2 \quad (2.1)$$

It follows that

$$\begin{aligned} J_T(w) &= \frac{1}{2} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \|y_i - y_j\|^2 \\ &= w^T \left[\frac{1}{2} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (x_i - x_j)(x_i - x_j)^T \right] w \end{aligned} \quad (2.2)$$

Let $S_T = \frac{1}{2} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (x_i - x_j)(x_i - x_j)^T$ and the mean vector $m = \frac{1}{N} \sum_{i=1}^N x_i$, then S_T can be calculated as follows:

$$S_T = \frac{1}{N} \sum_{i=1}^N (x_i - m)(x_i - m)^T \quad (2.3)$$

The projections $\{w_1, w_2, \dots, w_k\}$ comprise an orthogonal set of vectors representing the eigenvectors of S_T associated with the k largest eigenvalues, where normally $k < d$.

2.2.2 ICA

Since PCA only imposes independence up to the second order of the samples, some researchers have proposed several independent component analysis (ICA) algorithms to further improve the representation performance. The basic idea of ICA is to represent a set of random variables using basis functions, where the components are statistically independent or as independent as possible [151, 159]. Assume that there are m observed scalar variables $x_1, x_2, \dots, x_m$, which are linear combinations of n independent components $s_1, s_2, \dots, s_n$. The independent components are mutually statistically independent and with zero mean. We denote the observed variables x_i as a observed vector $X = (x_1, x_2, \dots, x_m)^T$ and the component variables s_i as a vector $S = (s_1, s_2, \dots, s_n)^T$. The relation between S and X can be modeled as

$$X = AS \quad (2.4)$$

where A is an unknown $m \times n$ full rank matrix, and each column of A represents a feature. The key point for ICA is the basic assumption that the component s_i are statistically independent, and many independence measures are used to seek the ICA feature space, such as kurtosis, neg-entropy, and mutual information.

2.2.3 LDA

Since PCA and ICA are unsupervised, they are more suitable for efficient data representation rather than classification. For many recognition tasks, supervised feature extraction methods are preferred because such supervised information can allow data to be better separated. LDA is one of the most typical supervised feature extraction methods. LDA aims to find a sets of projection axes such that the Fisher criterion (the ratio of the between-class scatter to the within-class scatter) is maximized after the projection. The between-class scatter S_B and the within-class scatter S_W are defined as [2]

$$S_B = \sum_{i=1}^c N_i (m_i - m)(m_i - m)^T \quad (2.5)$$

$$S_W = \sum_{i=1}^c \sum_{j=1}^{N_i} (x_{ij} - m_i)(x_{ij} - m_i)^T \quad (2.6)$$

where x_{ij} denotes the j th training sample of the i th class, m_i is the mean of the training sample of the i th class, and m is the mean of all the training samples. The Fisher criterion is defined as

$$\max_w J(w) = \frac{w^T S_B w}{w^T S_W w} \quad (2.7)$$

The corresponding projections $\{w_1, w_2, \dots, w_k\}$ comprise a set of the eigenvectors of the following generalized eigenvalue function

$$S_B w = \lambda S_W w \quad (2.8)$$

Let $\{w_1, w_2, \dots, w_k\}$ be the eigenvectors corresponding to the k largest eigenvalues $\{\lambda_i | i = 1, 2, \dots, k\}$ decreasingly ordered (i.e., $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$), then $W = [w_1, w_2, \dots, w_k]$ is the learned mapping of LDA. Since the rank of S_B is bounded by $c - 1$, k is at most equal to $c - 1$.

While LDA is an effective feature extraction method and has attained great success in many applications, it nevertheless suffers from two shortcomings: one is the total number of features extracted by LDA is limited to $c - 1$, which may be a bottleneck if the number of the classes is small, especially for a two-class recognition problem. Although it was demonstrated that LDA could outperform PCA in this particular case, it may not be the case for most of the two-class classification problems. Another shortcoming is that there is usually a small size sample (SSS) problem occurred in LDA. Some researchers have proposed several solutions to overcome these shortcomings including preprocessing using PCA [2] and regularizing of the within-class scatter [56–59].

2.2.4 LPP

Recently, there are considerable interest in manifold-based subspace learning for feature extraction [8–22, 24, 24–27]. The basic motivation of these methods is that high-dimensional face samples can be considered as a set of geometrically related points lying on or nearby a smooth, low-dimension manifold [5, 6]. The most representative such algorithm is locality preserving projections (LPP) [8]. The aim of LPP is to seek one low-dimensional subspace that preserves the intrinsic geometric structure of the original samples, allowing samples lying in a neighborhood to keep their locality relationship after projection.

LPP first constructs one affinity graph to characterize the neighborhood relationship of the training set and then seeks one low-dimensional embedding to preserve the intrinsic geometry and local structure. The objective function of LPP is formulated as follows [8]:

$$\min \sum_{ij} \|y_i - y_j\|^2 S_{ij} \quad (2.9)$$

where y_i and y_j are the low-dimensional representation of x_i and x_j . The affinity matrix S can be defined as:

$$S_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2/t), & \text{if } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ 0 & \text{otherwise} \end{cases} \quad (2.10)$$

where t and k are two dependent pre-specified parameters defining the local neighborhood, and $N_k(x)$ denotes the k nearest neighbors of x . Following some simple algebra

deduction steps [8], one can obtain

$$\frac{1}{2} \sum_{ij} \|y_i - y_j\|^2 S_{ij} = wXLX^T w \quad (2.11)$$

where D is a diagonal matrix and its entries are column sums of S , i.e., $D_{ii} = \sum_{ij} S_{ji}$, and $L = D - S$ is the Laplacian matrix. Matrix D provides a natural measure on the data points, and the bigger the value D_{ii} is, the more “important” y_i is. Usually, one can impose a constraint that:

$$y^T D y = 1 \Rightarrow wXDX^T w = 1. \quad (2.12)$$

Then, this minimization problem can be converted into solving the following constrained optimization problem:

$$\begin{aligned} w_{opt} &= \arg \min_w w^T X L X^T w \\ s.t. \quad & w^T X D X^T w = 1. \end{aligned} \quad (2.13)$$

Finally, the bases of LPP are the eigenvectors of the following generalized eigenvalue problem:

$$X L X^T w = \lambda X D X^T w \quad (2.14)$$

Let $w_1, w_2, \dots, w_k$ be the solutions of (2.14), ordered according to their eigenvalues, $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_k$, then $W = [w_1, w_2, \dots, w_k]$ be the resulted mapping of LPP.

2.2.5 DLPP

As LPP is original unsupervised, it may be suboptimal for recognition tasks. To enhance its discriminant power, Yu *et al.* [20] proposed a discriminant locality preserving projections (DLPP) method for image recognition. The aim of DLPP is to preserve the intrinsic geometry structure of original data and separate them in the low-dimensional subspace according to their labels. The objective function of DLPP is formulated as follows [20]:

$$\min \frac{\sum_{c=1}^C \sum_{i,j=1}^C \|y_i - y_j\|^2 W_{ij}^c}{\sum_{i,j=1}^C (m_i - m_j)^2 B_{ij}} \quad (2.15)$$

where C is the number of the classes, n_c is the number of samples in the c th class, y_i^c and y_j^c represent the i th and j th sample in the c th class, m_i and m_j are the mean of the i th

and j th class, i.e., $m_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_k^i$, $m_j = \frac{1}{n_j} \sum_{k=1}^{n_j} y_k^j$, where n_i and n_j are the number of samples in the i th and j th class, respectively, W_{ij}^c and B_{ij} are weight matrices defined as: $W_{ij}^c = \exp(-\|x_i^c - x_j^c\|^2/t_1)$, $B_{ij} = \exp(-\|f_i - f_j\|^2/t_2)$, $f_i = \frac{1}{n_i} \sum_{k=1}^{n_i} x_k^i$, t_1 and t_2 are two empirically pre-specified parameters.

By using the simple algebra deduction proposed in [20], the optimization problem can be reduced to

$$\min \frac{w^T X L X^T w}{w^T F H F^T w} \quad (2.16)$$

where $L = D - W$, $D_{ii} = \sum_j W_{ij}$, $D = \text{diag}\{D_1, D_2, \dots, D_C\}$, $W = \text{diag}\{W_1, W_2, \dots, W_C\}$, $H = E - B$, $E_{ii} = \sum_j B_{ij}$.

Finally, the bases of DLPP are the eigenvectors of the following generalized eigenvalue problem:

$$X L X^T w = \lambda F H F^T w \quad (2.17)$$

Let $w_1, w_2, \dots, w_k$ be the solutions of (2.17), ordered according to their eigenvalues, $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_k$, then $W = [w_1, w_2, \dots, w_k]$ is the mapping of DLPP.

2.2.6 MFA

Yan *et al.* [17] generalized some existing subspace learning methods and proposed a new discriminant manifold learning algorithm: margin fisher analysis (MFA). MFA adopts a graph embedding framework and designs two graphs: one intrinsic graph that characterizes the intraclass compactness and another penalty graph which characterizes the interclass separability. Let S_c and S_p be the measure of the intrinsic and penalty graph defined as

$$\begin{aligned} S_c &= \sum_{ij} \sum_{i \in N_{k_1}(j) \text{ or } i \in N_{k_1}(j)} \|w^T x_i - w^T x_j\|^2 \\ &= 2w^T (D^c - W^c) X^T w \end{aligned} \quad (2.18)$$

$$\begin{aligned} S_p &= \sum_{ij} \sum_{i \in P_{k_2}(j) \text{ or } i \in P_{k_2}(j)} \|w^T x_i - w^T x_j\|^2 \\ &= 2w^T (D^p - W^p) X^T w \end{aligned} \quad (2.19)$$

where

$$W_{ij}^c = \begin{cases} 1, & \text{if } x_i \in N_{k_1}(x_j) \text{ or } x_j \in N_{k_1}(x_i) \\ 0 & \text{otherwise} \end{cases} \quad (2.20)$$

and

$$W_{ij}^p = \begin{cases} 1, & \text{if } x_i \in P_{k_2}(x_j) \text{ or } x_j \in P_{k_2}(x_i) \\ 0 & \text{otherwise} \end{cases} \quad (2.21)$$

Here $N_{k_1}(x)$ denotes the k_1 nearest neighbors of x in the same class and $P_{k_2}(x)$ denotes the k_2 nearest neighbors of x in different classes, and k_1 and k_2 are two pre-determined parameters.

The projection of MFA can be sought by solving the following optimization problem:

$$w_{opt} = \arg \min_w \frac{w^T X(D^c - W^c)X^T w}{w^T X(D^p - W^p)X^T w} \quad (2.22)$$

Then, w can be obtained by solving a generalized eigenvalue function

$$Lw = \lambda Bw \quad (2.23)$$

where $L = X(D^c - W^c)X^T$ and $B = X(D^p - W^p)X^T$. Similar to LPP, let $w_1, w_2, \dots, w_k$ be the solutions of (2.23), ordered according to their eigenvalues, $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_k$, then $W = [w_1, w_2, \dots, w_k]$ is the resulted mapping of MFA.

2.2.7 NMF

There are psychological and physiological evidences that humans recognize objects through part-based representations in the brain [35]. Based on this belief, another type of subspace learning method called nonnegative matrix factorization (NMF) [35, 36] has been proposed in the literature to impose nonnegative constraints in the bases and representation coefficients, resulting in better representations of part-based face features such as mouth, nose and eyes, which have more discriminant information than other face parts such as cheek and forehead as indicated in [35, 36].

Given a nonnegative $d \times N$ matrix V , each column represents one biometrics sample, r is one rank parameter, the NMF model finds a nonnegative $d \times r$ matrix W and another nonnegative $r \times N$ matrix H such that their product WH approximates to V , i.e.,

$$V \approx WH = \sum_{k=1}^r W_{ik} H_{kj} \quad i = 1, 2, \dots, d, j = 1, 2, \dots, N. \quad (2.24)$$

Usually, r is chosen to be smaller than $\min\{d, N\}$ and is generally chosen such that $(d+N)r \leq dN$. In order to find a better approximate factorization, one objective function that minimizes the squared error between V and WH with nonnegative constraints can be defined as follows:

$$\begin{aligned} F(W, H) &= \|V - WH\|^2 = \sum_{i,j} (V_{i,j} - (WH)_{i,j})^2 \\ \text{s.t. } & W, H \geq 0 \end{aligned} \tag{2.25}$$

To solve this nonconvex optimization problem, the classical Expectation-Maximization (EM) can be applied to find the local minima.

As NMF is originally unsupervised, some recent attempts have exploited the discriminant information and formulated several discriminant nonnegative matrix factorization (DNMF) algorithms [38–40] to enhance the separability of the extracted part-based features. Further, to better characterize the sparseness of the bases and coefficients, more recent work has included sparse information into the formulation of NMF and derived many sparse-based NMF methods, such as local nonnegative matrix factorization (LNMF) [41], nonnegative matrix factorization with sparseness constraints (NMFSC) [42], and nonsmooth nonnegative matrix factorization (NsNMF) [43], to extract features which are sparser than NMF features. Some recent methods introduced geometrical information into the objective function of NMF and formulated geometrically motivated NMF extensions, such as topology preserving nonnegative matrix factorization (TPNMF) [44], nonnegative graph embedding (NGE) [45], and graph regularized nonnegative matrix factorization (GNMF) [46], to preserve the nonlinear manifold structure of face samples for extracting part-based features. More recently, NMF was extended to reproducing kernel Hilbert space (RKHS) and a polynomial kernel nonnegative matrix factorization (PNMF) [47] algorithm was formulated to extract nonlinear nonnegative features in high-dimensional feature space. All these NMF extensions can enhance the representation capability and the recognition performance, and a good recent survey of NMF and its extensions can be found in [48].

In the following Chapters 3-7, we present our proposed enhanced linear subspace algorithms to address the limitations of existing linear subspace approaches.

Chapter 3

Doubly Weighted Subspace Learning

3.1 Introduction

Existing subspace learning-based methods consider each face sample holistically and ignore the fact that different face parts may have different discriminative power for face representation and recognition. Recent studies have provided preliminary and convincing evidence of this fact [49–52]. Hence, it is more reasonable to assign different weights to different face parts for face representation. Moreover, most existing subspace learning methods consider each sample in the training set contributing equally in the training phase, i.e., the contribution of each sample is considered to be $1/N$, where N is the number of samples in the training set. However, this may not always be the case in real-world applications. For example, when face samples for the training and recognition phases are acquired under different environments or using different sensing methods (still versus video cameras), the face samples acquired may have different correlations or underlying distributions. Moreover, it is also common to have different number of useful training samples due to pose and illumination variations. Hence, an efficient face recognition algorithm that can deal with imbalanced training samples and place proper emphasis on different face samples and parts is of great interest and significance.

Motivated by the above observations and reasons, we propose in this paper a doubly weighted subspace learning approach for face recognition, imposing larger weights on face samples which have low similarity to others in the same class and applying different emphases on different face parts according to their importances for face recognition. We first construct two weighting matrices based on pairwise similarity of face samples

within a same class and discriminant score of each pixel within a face sample to duly emphasize the between-sample and within-sample features. We then incorporate these two weighting matrices into three conventional subspace learning methods, namely PCA, LDA and NMF. Being able to weight the face samples and parts differently allows the proposed approach to deal with imbalanced training samples more effectively.

The remaining of this chapter is organized as follows. Section 3.2 details the proposed doubly weighted approach and its incorporation with existing subspace learning algorithms for face representation and recognition. Section 3.3 demonstrates the effectiveness of the proposed approach through a series of face representation and recognition experiments. Finally, Section 3.4 provides our concluding remarks.

3.2 Doubly Weighted Subspace Learning

The proposed doubly weighted approach makes use of two weighting matrices to depict the discriminant and geometrical information of face features. The first matrix, referred to as between-sample weighting matrix A_b , imposes different weights on different face samples to reflect their respective importances in the training phase. The other matrix, referred to as within-sample weighting matrix A_w , gives appropriate weights to pixels at different face parts. As both between-sample and within-sample features are weighted simultaneously, we therefore name our proposed method a “doubly weighted” approach.

3.2.1 Motivation

Most existing subspace learning methods consider each sample in the training set contributing equally in the training phase, i.e., the contribution of each sample is usually set to be $1/N$, where N is the number of the training samples. However, this assumption of equal contribution does not necessarily hold in real-world applications. For example, when face images are sampled from a video sequence during the training phase, it is possible to have many similar face samples which may not necessarily reflect the distribution of the faces acquired in the testing (recognition) phase. Hence, it may be more reasonable to assign different weight to each face sample to reflect its different importance during the training. Moreover, in many practical recognition applications, it is difficult and

expensive to collect enough training samples, and we need to exploit the discriminant information from limited number of training samples.

In other words, most existing subspace learning methods do not consider the different importance of each sample in different classes. These methods may work well when the testing samples are drawn from the same underlying distribution as the training samples. Unfortunately, due to the potential differences between the image acquisitions in the training and testing phases (lightings, backgrounds, sensors, etc.), it is difficult to ascertain that the training and testing samples are drawn from the same underlying distribution. Hence, the different representative capability of each sample in the training set should be fully exploited to make the learned subspace to be more generic in dealing with the variants of the testing samples. We provide below an extreme example to illustrate this point.

Suppose there are 5 training samples within a class of the training set, denoted as S_1 , S_2 , S_3 , S_4 and S_5 , respectively. Assume that samples S_1 , S_2 , S_3 and S_4 are identical and different from sample S_5 . If the training samples do not share the same distribution as the testing samples, the contribution (per sample) of S_1 , S_2 , S_3 and S_4 are apparently smaller than that of S_5 for learning a representative subspace. Under this extreme scenario, a more reasonable choice is to assign different weights to these samples to reflect their different importances despite they are from the same class.

Moreover, in real world applications of face recognition, it is common to have different number of useful training samples due to pose and illumination variations. Hence, an efficient face representation and recognition algorithm that can handle imbalanced training samples and place proper emphases on different face samples is of great interest and significance.

One possible solution to this problem is to make each training sample to be as “representative” as possible. For example, we can define its representative power (RP) of each training sample x_{ij} in the i th class as follows

$$RP(x_{ij}) = 1 - \max_{r, r \neq j} S(x_{ij}, x_{ir}) \quad (3.1)$$

where $S(x, y)$ denotes a similarity measure of two samples x and y , and $0 \leq S(x, y) \leq 1$.

According to this definition, we can see that different samples in each class are assigned with different weights to depict their differences from other samples in the same class. Generally, the larger “representative power” of a training sample, the larger weight should be assigned to it.

On the other hand, it is likely that different pixels in different spatial positions of each face sample may have different discriminative power. Most existing subspace learning methods extract features based on some statistical analysis and do not consider the physical meaning of the features such as the eyes and mouths of the face samples. That is, each element in a face image is treated equally and acts the same in subspace learning. However, due to the different spatial importance of each pixel in a face image, the pixel with different values and spatial locations could be assigned different weights for discrimination. For example, pixels at important facial features such as eyes, mouth, nose, etc., could be emphasized and others such as cheek and forehead could be deemphasized.

Motivated by the above-mentioned reasons, we hence propose in this paper a doubly weighted approach to explicitly exploit the different importance of each sample and each pixel of a sample simultaneously.

3.2.2 Between-Sample Weighting Matrix

To better characterize the different importance of each sample in the training set, a proper weighting matrix can be used to better reflect their different “representative” power in subspace learning. One reasonable rule could be that “the higher the similarity between two samples in a same class, the smaller the weight should be given to them”. Fig. 3.1 provides one toy example to illustrate this rule. The figure shows three classes of samples denoted by squares, circles and triangles, and each class has different number of samples. It can be seen that the number of samples in the circle or square class is fewer than that in the triangle class, and the samples in the triangle class have a higher correlation among them. Hence, smaller weights should be assigned to samples in the triangle class. Furthermore, different weights should also be assigned to the samples in the same class. For example, as the distance (d_{13}) between sample S_1 and sample S_3 in the circle class is much smaller than that (d_{12}) of S_1 and S_2 , smaller weights should be given to S_1 and S_3 compared with that of S_1 and S_2 . To effectively characterize the

Fig. 3.1: An example showing the similarities and differences among samples in each class.

similarity and structure of the samples, we devise the following algorithm to construct the between-sample weighting matrix.

Given a set of properly normalized $w \times h$ face images, we construct one training set of column image vectors X_{cj} , where $X_{cj} \in \mathbb{R}^{d=wh}$, by lexicographically ordering the pixels of the j th image of the c th subject. Assume that there are n different subjects and the c th subject has n_c samples. Our algorithm of constructing A_b is as follows:

- 1) For each sample in the c th class, $c = 1, 2, \dots, n$, calculate the similarity between each pair of samples in the class:

$$s_{i,j}^c = S(X_{ci}, X_{cj}), i \neq j, 1 \leq i, j \leq n_c \quad (3.2)$$

where the similarity metric is chosen as

$$S(X_{ci}, X_{cj}) = \exp^{-\|X_{ci} - X_{cj}\|^2 / \sigma} \quad (3.3)$$

with a value ranging from 0 to 1, and σ is a suitable constant.

- 2) Find the maximum similarity value S_M for X_{cj} among all pairwise similarities between X_{cj} and other samples in the c th class,

$$S_M(X_{cj}, :) = \max\{s_{i,j}^c\} \quad (3.4)$$

where $j = 1, 2, \dots, n_c$.

3) Assign to sample X_{cj} a weight a_{cj} defined as

$$a_{cj} = 1 - S_M(X_{cj}, :) \quad (3.5)$$

4) Normalize a_{cj} to make the weights sum to unity, i.e., $\sum_{c=1}^{c=n} \sum_{j=1}^{n_c} a_{cj} = 1$:

$$a_{cj} = \frac{a_{cj}}{\sum_{c=1}^n \sum_{j=1}^{n_c} a_{cj}} \quad (3.6)$$

5) Construct an $N \times N$ diagonal weighting matrix A_b with diagonal element a_{cj} , where $N = \sum_{c=1}^n n_c$ and

$$A_b = \text{diag}(a_{11}, a_{12}, \dots, a_{1n_1}, a_{21}, \dots, a_{nn_n}).$$

3.2.3 Within-Sample Weighting Matrix

With disparate influences on face recognition, different parts of face features should be assigned with different weights to attain better discrimination performance. Intuitively, some face parts such as eyes, nose and mouth may have more discriminative power than other parts such as cheek and forehead. We propose as follows a pixel-region fusion weighting algorithm to exploit the discriminant information of different face parts:

1) Apply a feature selection scheme based on a manifold-based maximum margin (MMM) criterion [10,12], described below, to calculate the discriminant capability of each pixel in the face samples.

Let f_{ri} denote the r th feature of the i th sample x_i , where $i = 1, 2, \dots, N$ and $r = 1, 2, \dots, d$. Let L_r denote the MMM score of the r th feature. We first construct a graph G to describe the locality structure of the samples. For each sample x_i , we find its k nearest neighbors and put an edge between x_i and its neighbors. Let $N(x_i) = \{x_i^1, \dots, x_i^k\}$ be the set of the k nearest neighbors of x_i . It follows that G can be defined as follows:

$$G_{ij} = \begin{cases} 1 & \text{if } x_i \in N(x_j) \text{ or } x_j \in N(x_i) \\ 0 & \text{otherwise} \end{cases} \quad (3.7)$$

To better characterize the locality and discriminant information of the samples, we split $N(x_i)$ into two subsets: $N_b(x_i)$ and $N_w(x_i)$, such that $N_b(x_i) \cap N_w(x_i) = \emptyset$ and $N_b(x_i) \cup N_w(x_i) = N(x_i)$, where $N_w(x_i)$ contains the neighbors with their

Fig. 3.2: Partitioning a face image into patches.

labels the same as that of x_i , and $N_b(x_i)$ contains the neighbors with different labels. We then derive two matrices G^c and G^p as follows:

$$G_{ij}^c = \begin{cases} \rho & \text{if } x_i \in N(x_j) \text{ or } x_j \in N(x_i) \text{ and } l(x_i) = l(x_j) \\ 0 & \text{otherwise} \end{cases} \quad (3.8)$$

and

$$G_{ij}^p = \begin{cases} 1 & \text{if } x_i \in N(x_j) \text{ or } x_j \in N(x_i) \text{ and } l(x_i) \neq l(x_j) \\ 0 & \text{otherwise} \end{cases} \quad (3.9)$$

where $l(x_i)$ and $l(x_j)$ denote the class label of x_i and x_j , and the values of k and ρ are empirically set to be 10 and 50 in our experiments, respectively.

- 2) Define two diagonal matrices D^c and D^p , where $D_{ii}^c = \sum_j G_{ij}^c$ and $D_{ii}^p = \sum_j G_{ij}^p$, and obtain the MMM matrices as $H^c = D^c - G^c$ and $H^p = D^p - G^p$, respectively. Compute the pixel-based MMM score for the r th feature as

$$L_r = \frac{f_r^T H^c f_r}{f_r^T H^p f_r} \quad (3.10)$$

where $r = 1, 2, \dots, d$.

- 3) For each face image x_i of $w \times h$ pixels, we partition it into P non-overlapping patches $x_i = (x_i^1, x_i^2, \dots, x_i^P)$ with the same size of $k \times l$ pixels as illustrated in Fig. 3.2. Let $m_q = (m_q^1, m_q^2, \dots, m_q^N)^T$ be the average of the q th patch of all the samples. Apply the similar approach to calculate the patch-based MMM score for each patch q as

$$R_q = \frac{m_q^T H^c m_q}{m_q^T H^p m_q} \quad (3.11)$$

where $q = 1, 2, \dots, P$.

Fig. 3.3: Visualization of a within-sample weighting matrix.

- 4) For each position (i, j) in a face sample, we find its pixel-based MMM score L_{ij} and patch-based score R_{ij} , where $L_{ij} = L(x)$, $R_{ij} = R(y)$, $x = (i - 1) \times w + j$, and $y = ([(i - 1) / k] + 1) \times k + ([(j - 1) / l] + 1)$. The operator $[c]$ retains the integer part of c .
- 5) For each pixel $f(i, j)$ in a face sample, we combine their pixel-based and patch-based MMM scores to obtain a fused score as follows:

$$t_{i,j} = L_{ij}R_{ij} \quad (3.12)$$

where $1 \leq i \leq w$ and $1 \leq j \leq h$.

- 6) Construct a $wh \times wh$ matrix A_w with diagonal element t_{ij} 's, i.e.,

$$A_w = \text{diag}(t_{1,1}, t_{2,1}, \dots, t_{w,1}, \dots, t_{w,h}).$$

Let $T(i, j) = t_{ij}$. We first normalize each element in T to the range of $[0, 255]$ as follows

$$T(i, j) = \left(\frac{255}{\max(T) - \min(T)} \right) T(i, j) \quad (3.13)$$

where $\max(T)$ and $\min(T)$ denote the maximum and minimum weight of T , respectively. Fig. 3.3 shows a within-sample weighting matrix of the PIE database, calculated from a training set containing 68 subjects and each subject has 10 different images. We can see from this figure that the within-sample weighting matrix emphasizes different importance of each pixel of the face image. Intuitively speaking, some parts on the face such as eyes and eyebrows are assigned with large weights while other regions are assigned with small weights.

We now briefly analyze the computational complexity of calculating the within-sample weighting matrix, which involves four steps: 1) constructing the graph G ; 2) computing the pixel-based MMM score; 3) computing the patch-based MMM score; and 4)

Fig. 3.4: Flowcharts of proposed doubly weighted approach.

computing the fused MMM score. These four steps require $O(dN^2)$, $O(dN^2)$, $O(\frac{d}{kl}N^2)$ and $O(d)$ operations, respectively. Hence, the overall computational complexity of the within-sample weighting matrix is in the order of $O(dN^2)$.

As PCA is more efficient than LDA and NMF, we select PCA and compare its computational complexity with that of the within-sample weighting matrix. The most efficient PCA algorithm requires $O(dN^2)$ operations [55], the same as that of our proposed within-sample weighting matrix. Hence, we can see that the within-sample weighting matrix does not notably increase the computational complexity of existing subspace learning methods.

3.2.4 Proposed Algorithms

We now present the proposed doubly weighted algorithms, referred to as DWPCA, DWLDA and DWNMF, by integrating the between-sample and within-sample weighting matrices into the existing PCA, LDA and NMF methods, respectively. The algorithms' flowcharts are depicted in Fig. 3.4.

- 1) Given N training samples X_{ij} containing n different subjects and each subject has n_i samples of $w \times h$ pixels, organize the training samples into an $wh \times N$ matrix

X , in which each column of X is formed by one training sample.

- 2) Construct the $N \times N$ between-sample weighting matrix A_b for X . Multiply X with A_b to obtain V' , i.e.,

$$X' = X A_b \quad (3.14)$$

- 3) Perform PCA, LDA and NMF for X' to obtain the $wh \times k$ bases W_{PCA} , W_{LDA} and W_{NMF} .
- 4) Construct the $wh \times wh$ within-sample weighting matrix A_w and then multiply A_w with W to obtain the new bases W_{PCA} , W_{LDA} , and W_{NMF} :

$$W_{DWPCA} = A_w W_{PCA} \quad (3.15)$$

$$W_{DWLDA} = A_w W_{LDA} \quad (3.16)$$

$$W_{DWNMF} = A_w W_{NMF} \quad (3.17)$$

- 5) Project the training and testing samples onto W_{DWPCA} , W_{DWLDA} , and W_{DWNMF} to extract the associated feature representations, respectively.

3.2.5 Relation with Existing Weighted Techniques

While weighted subspace learning has not received much attention in the domain of machine learning and pattern recognition, there have been a few attempts on weighting-based object recognition in the literature [49, 50, 53, 53, 54]. For example, motivated by that the eigenfaces derived from the training face samples can well characterize the variation of different faces at different positions, Xie *et al.* [49] and Lin *et al.* [50] applied the first eigenface, the one that corresponds to the largest eigenvalue of the covariance matrix derived from the training set, to weight the importance of each pixel of a face image. However, such weighting only exploits the variation of the training samples, which may not be equivalent to the discriminant information and is arguably not optimal for recognition purpose. Tan *et al.* [51] used rectangular weighting functions to characterize the different discriminant information in different parts of a face sample, and assign different weights to them, e.g., 1 to the eyes and mouth parts, 0.5 to other facial parts, and 0 to the background. However, this weighting is rather ac-hoc and unlikely can fully characterize the discriminant information of a face sample. Furthermore, as both

face pixels and regions could provide different discriminant information useful for face feature extraction and recognition, a suitable weighting function should consider these two aspects, simultaneously.

In addition, Guillaumet *et al.* [53,54] introduced a weighted technique to assign different weights to different training samples according to their occurrence probabilities in the training set. However, even different weights were assigned to different samples and the redundancy among the bases were reduced to a certain extent, the pairwise similarity of the samples in the same class was not considered and thus the geometrical information of the original training samples might not be fully exploited. As illustrated in Fig. 3.1, different weights should also be assigned to the samples even they are in the same class, as their representative power may be different from one to another. Furthermore, within-sample weighting was not considered in [53,54], and the weighting method can be regarded as a special case of our proposed doubly weighted approach when the pairwise similarity of the samples in the same class is uniform and each pixel and each part in the face contributes equally to face feature extraction and recognition.

3.3 Experiments and Results

In this section, we evaluate the effectiveness of our proposed doubly weighted algorithms for face representation and recognition.

3.3.1 Data Preparation

We have used the Extended Yale B [104,105] and CMU PIE [106] face databases in our experiments. The Extended Yale B database consists of 2432 gray images of 38 subjects with 64 different illumination conditions. The PIE database comprises 68 subjects with 41368 face images of different poses, illuminations and expressions. To limit our evaluation on frontal face recognition, we selected a subset of the PIE database consisting of 3060 frontal face images of different expressions and illuminations from 68 subjects with 45 images from each subject. Each image was manually aligned according to the eyes' positions and resized into 32×32 pixels. Fig. 3.5 shows some aligned and resized face samples from the two databases.

Fig. 3.5: Aligned and resized face samples of one subject in the (a) Extended Yale B database and (b) PIE database.

We designed several sets of experiments with balanced or imbalanced training sets. For the Extended Yale B database, we randomly divided the 38 subjects into four non-overlapping groups (G1, G2, G3 and G4). The first two groups have 10 subjects each and the remaining two groups have 9 subjects each. Similarly, we divided the PIE database into four non-overlapping groups (G1, G2, G3 and G4) of 17 subjects each. We then randomly selected m_1 , m_2 , m_3 and m_4 samples, respectively, from each subject in the four groups of the two databases to form the training set, and used the remaining samples for testing. When the numbers of m_1 , m_2 , m_3 and m_4 are equal to each other, the training set is balanced; otherwise, it's imbalanced.

3.3.2 Face Representation

We set $(m_1, m_2, m_3, m_4) = (10, 10, 20, 20)$ for the PIE database and conducted face representation experiments to visualize the features extracted by PCA, LDA, NMF and their doubly weighted variants. Fig. 3.6 shows the first 10 bases of PCA, LDA, NMF, DWPCA, DWLDA and DWNMF, respectively.

To quantify the redundancy of the bases, we applied pairwise mutual information to calculate the correlation of two bases. Let w_1 and w_2 be two bases with size of $w \times h$. Their mutual information can be calculated as follows:

$$I(w_1, w_2) = H(w_1) + H(w_2) - H(w_1, w_2) \quad (3.18)$$

where $H(X) = -\sum_j p_j \log p_k$ is the entropy of X , p denotes the probability mass function of X , $H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$ is the joint entropy of X and Y , respectively. Table 3.1 tabulates the average pairwise mutual information (APMI) of the bases learned from the six subspace learning methods. We can see from Table 3.1 that the average

Fig. 3.6: Set of the first 10 basis images of (a) PCA, (b) LDA, (c) NMF, (d) DWPCA, (e) DWLDA, and (f) DWNMF methods.

Table 3.1: Average pairwise mutual information (APMI) of six subspace learning methods.

Method	PCA	LDA	NMF	DWPCA	DWLDA	DWNMF
APMI	2.7129	2.3794	1.7382	1.6160	1.2750	1.6727

pairwise mutual information of DWPCA, DWLDA and DWNMF is less than that of PCA, LDAF and NMF by 1.0969, 1.1044 and 0.0655, respectively. The results show that the proposed doubly weighted approach can effectively reduce the redundancy of the bases as compared to the original methods.

According to the above results, we have made the following observations:

- 1) The bases of doubly weighted subspace learning algorithms are intrinsically different from those of conventional non-weighted methods. More specifically, the bases of doubly weighted subspace learning algorithms contain more salient meanings because important elements of these bases have been emphasized and some others have been deemphasized. The reason is that the doubly weighted algorithms weight the bases by using a discriminative within-sample weighting matrix, which results in more salient features in these bases.

Fig. 3.7: Normalized kurtosis of NMF and DWNMF.

- 2) The difference of the bases computed from the non-weighted, unsupervised learning methods (such as PCA and NMF) and the corresponding doubly weighted variants (such as DWPCA and DWNMF) are larger than that of the bases computed from the supervised methods (such as LDA and DWLDA). This is because the bases of the supervised learning method such as LDA originally contain some discriminant information and those of the unsupervised learning methods such as PCA and NMF contain little discriminant information. Hence, the discriminative weighting operation has more effect on unsupervised learning methods rather than supervised methods.
- 3) The bases of DWNMF are much sparser than that of NMF. This implies that the proposed doubly weighted approach can enhance the sparseness of NMF.

To quantify the degree of sparseness of each basis (w) of NMF and DWNMF, we use the normalized kurtosis as a sparseness measure criterion, which was suggested in [38] and defined as follows:

$$\kappa(w) = \frac{\sum_i (w_i - \bar{w})^4}{(\sum_i (w_i - \bar{w})^2)^2} - 3 \quad (3.19)$$

where $w = [w_1, w_2, \dots, w_d]^T$ and $\bar{w} = 1/d \sum_{i=1}^d w_i$. Fig. 3.7 reveals the average sparseness of NMF and DWNMF versus different number of bases. We can observe that the average sparseness of the bases of DWNMF is consistently higher than that of NMF.

Table 3.2: Number of samples in the four training sets from the Extended Yale B and PIE Databases.

	Extended Yale B		PIE	
	Exp. 1	Exp. 2	Exp. 1	Exp. 2
Balanced database (BD)	10, 10, 10, 10	20, 20, 20, 20	5, 5, 5, 5	10, 10, 10, 10
Imbalanced database (ID)	10, 15, 20, 25	15, 20, 25, 30	5, 5, 10, 10	10, 10, 20, 20

3.3.3 Face Recognition

In this subsection, we present our face recognition experiments conducted with several sets of m_1 , m_2 , m_3 and m_4 training images from the Extended Yale B and PIE databases as indicated in Table 3.2. We compared our proposed doubly weighed subspace learning algorithms with other existing non-weighted methods in terms of recognition accuracy. We applied the nearest neighbor rule in the Euclidean space to perform recognition. Note that the recognition accuracy reported here is the rank-1 recognition rate of each experiment. It is defined as N_c/N_t , where N_c is the number of correctly recognized testing samples and N_t is the number of total testing images. To reduce the variation in the recognition results, we computed the average recognition accuracy (ARA) of 20 runs in our experiments.

As the advantage of the proposed doubly weighted approach stems from two different weightings: between-sample weighting and within-sample weighting, we also evaluated the performance when only one of the weightings is applied to reveal their respective effects. Let W1 and W2 denote the between-sample weighting and within-sample weighting, respectively. We thus formulated six singly weighted NMF-based algorithms, i.e., W1PCA, W2PCA, W1LDA, W2LDA, W1NMF, and W2NMF. Further, we designed several training sets with balanced and imbalanced samples for face recognition experiments to show the superiority of the proposed methods. We compared our proposed doubly methods with the conventional non-weighted subspace learning algorithms including PCA [1], LDA [2] and NMF [36, 37]. We report here the best result of each algorithm under comparison by exploring all possible feature dimensions. The average recognition accuracy and standard deviation are tabulated in Tables 3.3-3.4, respectively. As can be seen, the proposed doubly weighted subspace learning methods always achieve higher recognition accuracy than existing non-weighted methods.

Table 3.3: Recognition performance comparison on the Extended Yale B database (mean $\pm$ std).

Method	Exp. 1 (BD)		Exp. 2 (BD)		Exp. 1 (ID)		Exp. 2 (ID)	
	Accu. (%)	Dim.	Accu. (%)	Dim.	Accu. (%)	Dim.	Accu. (%)	Dim.
PCA	42.35 $\pm$ 2.9	340	56.76 $\pm$ 1.2	750	37.49 $\pm$ 2.3	750	44.24 $\pm$ 2.6	750
W1PCA	43.36 $\pm$ 2.8	340	57.34 $\pm$ 1.4	750	38.23 $\pm$ 2.6	750	44.98 $\pm$ 2.7	750
W2PCA	43.62 $\pm$ 2.7	340	57.68 $\pm$ 1.5	750	38.46 $\pm$ 2.5	750	45.69 $\pm$ 2.8	750
DWPCA	44.68 $\pm$ 2.9	340	58.98 $\pm$ 1.2	750	40.32 $\pm$ 2.4	750	46.64 $\pm$ 2.6	750
LDA	68.56 $\pm$ 1.7	37	78.11 $\pm$ 2.4	37	58.29 $\pm$ 3.1	37	59.86 $\pm$ 1.4	37
W1LDA	70.32 $\pm$ 1.7	37	80.22 $\pm$ 2.6	37	60.29 $\pm$ 3.1	37	61.84 $\pm$ 1.5	37
W2LDA	70.65 $\pm$ 1.9	37	80.44 $\pm$ 2.5	37	61.21 $\pm$ 3.3	37	61.34 $\pm$ 1.6	37
DWLDA	72.44 $\pm$ 2.0	37	82.32 $\pm$ 2.6	37	62.56 $\pm$ 3.2	37	63.24 $\pm$ 1.5	37
NMF	63.15 $\pm$ 1.2	180	82.95 $\pm$ 2.5	220	62.18 $\pm$ 1.4	320	59.74 $\pm$ 2.1	340
W1NMF	64.40 $\pm$ 2.4	220	83.13 $\pm$ 1.3	220	62.97 $\pm$ 1.8	320	61.75 $\pm$ 2.7	340
W2NMF	65.59 $\pm$ 2.4	200	83.38 $\pm$ 1.6	220	63.14 $\pm$ 3.9	320	60.03 $\pm$ 4.1	340
DWNMF	75.51 $\pm$ 1.6	220	84.08 $\pm$ 2.0	240	74.18 $\pm$ 1.3	320	63.55 $\pm$ 2.3	320

Table 3.4: Recognition performance comparison on the PIE database (mean $\pm$ std).

Method	Exp. 1 (BD)		Exp. 2 (BD)		Exp. 1 (ID)		Exp. 2 (ID)	
	Accu. (%)	Dim.	Accu. (%)	Dim.	Accu. (%)	Dim.	Accu. (%)	Dim.
PCA	68.63 $\pm$ 3.4	140	83.05 $\pm$ 2.7	240	61.22 $\pm$ 3.5	160	70.06 $\pm$ 3.7	260
W1PCA	69.43 $\pm$ 3.2	140	83.72 $\pm$ 2.5	240	62.04 $\pm$ 3.3	160	70.38 $\pm$ 3.4	260
W2PCA	68.96 $\pm$ 3.0	140	83.58 $\pm$ 2.6	220	61.89 $\pm$ 3.2	160	70.26 $\pm$ 2.8	260
DWPCA	70.02 $\pm$ 2.8	140	83.86 $\pm$ 2.4	220	62.46 $\pm$ 3.0	140	70.64 $\pm$ 2.6	240
LDA	79.08 $\pm$ 2.8	67	86.32 $\pm$ 2.4	67	70.04 $\pm$ 3.1	67	74.07 $\pm$ 3.4	67
W1LDA	79.68 $\pm$ 2.4	67	86.78 $\pm$ 2.3	67	70.68 $\pm$ 3.0	67	74.54 $\pm$ 3.2	67
W2LDA	79.88 $\pm$ 2.2	67	86.92 $\pm$ 2.2	67	70.46 $\pm$ 3.1	67	74.26 $\pm$ 3.0	67
DWLDA	80.24 $\pm$ 2.3	67	87.48 $\pm$ 2.3	67	71.42 $\pm$ 3.0	67	75.12 $\pm$ 2.8	67
NMF	80.01 $\pm$ 2.7	100	90.03 $\pm$ 3.3	160	70.23 $\pm$ 2.6	120	76.04 $\pm$ 2.8	160
W1NMF	80.40 $\pm$ 2.8	100	90.24 $\pm$ 3.4	160	72.07 $\pm$ 3.5	120	76.32 $\pm$ 3.8	160
W2NMF	80.21 $\pm$ 1.8	120	90.09 $\pm$ 1.7	160	70.80 $\pm$ 3.2	120	76.04 $\pm$ 3.5	160
DWNMF	84.41 $\pm$ 2.3	100	90.79 $\pm$ 1.4	140	73.52 $\pm$ 2.0	100	78.97 $\pm$ 2.1	140

We also evaluated the effect of different region sizes on the performance of the proposed doubly weighted algorithms. We partitioned the training images and testing images into various region sizes and repeated the imbalanced face recognition experiments on the PIE face database with $(m_1, m_2, m_3, m_4) = (5, 5, 10, 10)$. Without loss of generality, we report in Fig. 3.8 the results for the cases of $k = l$ for illustration purposes. As can be seen, when the region size is 32×32 , the region covers the entire face and hence the discriminant power of region diminishes. Similarly, when the region size is 1×1 , the discriminant score of each region is equal to that of each pixel, and hence the discriminant

Fig. 3.8: Performance of proposed DWPCA, DWLDA and DWNMF methods versus different region sizes.

power of each region also decreases. We can observe from the results that the optimal region size is 4×4 in our experiments.

Lastly, we compared the performance of the proposed doubly weighted methods with two existing weighted techniques: one uses the eigenface corresponding to the largest eigenvalue of the covariance matrix computed from the training set to reflect the importance of different parts in the face [49], and another weights the importances of different training samples by their occurrence possibilities [53]. As DNMF had performed the best among all the doubly weighted methods, we applied these two weighted techniques to DNMF and compared their performance with our proposed W1DNMF, W2DNMF and DWDNMF algorithms. Table 3.5 tabulates the results on the Extended Yale B database, in which our proposed weighted approach achieve higher recognition accuracy than the two weighted techniques compared. The is mainly due to that our weighted approach is supervised and more discriminant information can be exploited in the proposed weighting matrix for feature extraction. Moreover, our proposed weighted approach has also considered the different importance of each sample in a same class and retained the geometrical information for feature extraction.

Table 3.5: Comparison with existing weighted methods on the Extended Yale B database. (mean $\pm$ std).

Method	Exp. 1 (BD)		Exp. 2 (BD)		Exp. 1 (ID)		Exp. 2 (ID)	
	Accu. (%)	Dim.	Accu. (%)	Dim.	Accu. (%)	Dim.	Accu. (%)	Dim.
Guillamet [53]	70.13 $\pm$ 2.1	200	78.48 $\pm$ 2.0	220	60.68 $\pm$ 3.6	320	62.98 $\pm$ 2.9	320
Xie [49]	71.67 $\pm$ 2.3	200	82.68 $\pm$ 1.9	200	61.96 $\pm$ 2.1	320	63.84 $\pm$ 1.6	320
W1DNMF	72.27 $\pm$ 1.7	200	83.52 $\pm$ 2.0	220	62.89 $\pm$ 3.7	320	64.71 $\pm$ 2.5	320
W2DNMF	75.09 $\pm$ 2.3	200	84.23 $\pm$ 1.7	200	63.09 $\pm$ 1.9	320	64.67 $\pm$ 1.4	320
DWDNMF	81.21 $\pm$ 2.6	180	87.69 $\pm$ 2.3	200	75.43 $\pm$ 2.8	300	79.11 $\pm$ 2.4	300

3.3.4 Discussion

The following observations can be made from the experimental results listed in Tables 3.1-3.5 and Fig. 3.8:

- 1) The proposed doubly weighted subspace learning methods notably achieve higher recognition accuracy than the corresponding non-weighted methods. The improvement in recognition is due to that the proposed doubly weighted approach can better emphasize more important parts in faces and deemphasize the less important parts, and also exploit both discriminant and geometrical information for face feature extraction.
- 2) The doubly weighted approach can achieve more recognition accuracy gains for NMF than the other subspace learning methods such as PCA and LDA. One possible reason for this is that NMF is a part-based feature extraction and PCA and LDA are holistic-based feature extraction methods, and our proposed doubly weighted approach emphasizes more important features and deemphasize the less important features in the feature extraction process. Hence, it can better exploit the important face parts for NMF and induce more improvements for NMF rather than PCA and LDA.
- 3) Both between-sample weighting and within-sample weighting can enhance the recognition performance. Furthermore, when the recognition is conducted for balanced training samples, the within-sample weighting is more important, but for imbalanced training samples, the between-sample weighting becomes more important. One possible reason for this is that an imbalanced training set can likely induce more redundancy, and properly weighting the training samples can effectively reduce the redundancy.

3.4 Chapter Summary

We have proposed in this chapter a doubly weighted approach for subspace learning based face representation and recognition. Considering that some face samples and parts are more effectual in representing and recognizing faces, we have constructed two weighting matrices based on pairwise similarity of face samples within a same class and discriminant score of each pixel within a face sample to duly exploit both the between-sample and within-sample features. We have incorporated these two weighting matrices into three conventional subspace learning methods, and obtained more discriminative face features for recognition. To the best of our knowledge, our approach is the first attempt that formally addresses the face representation and recognition problem by explicitly and simultaneously considering the between-sample and within-sample weightings for face feature extraction. Experimental results on two public benchmark face databases have demonstrated the effectiveness and superiority of our proposed doubly weighted approach. Partial work presented in this chapter was published in [161] and [162].

Chapter 4

Parametric Regularized Subspace Learning

4.1 Introduction

Recently, a number of manifold learning algorithms have been proposed to discover the geometric property of high-dimensional feature spaces, and they have been successfully applied to face recognition. The most representative such algorithm is locality preserving projections (LPP) [8]. As LPP is originally unsupervised, more recent work has included supervised information into the formulation and derived many discriminant-based LPP algorithms, such as discriminant locality preserving projections (DLPP) [20], orthogonal discriminant locality preserving projections (ODLPP) [21], and uncorrelated discriminant locality preserving projections (UDLPP) [22], to enhance the recognition performance. Despite some differences in their objective functions, these discriminant extensions adopt a similar formulation and lead to the following generalized eigenvalue equation:

$$S_1 w = \lambda S_2 w \quad (4.1)$$

where S_1 and S_2 are two matrices to be minimized and maximized, respectively. Table 4.1 shows these matrices for LPP and its three discriminant extensions. (The definitions of matrices L , D , H , W , and G can be deduced from the respective papers [20,21,22].) In fact, S_2 is usually singular in face recognition, which stems from the fact that the number of training images is usually much smaller than the dimension of each image, a deficiency that is generally known as SSS problem.

Table 4.1: S_1 and S_2 for LPP and its three discriminant extensions.

Method	S_1	S_2
LPP [8]	XLX^T	$XD X^T$
DLPP [20]	XLX^T	FHF^T
ODLPP [21]	XLX^T	FHF^T
UDLPP [22]	$X(L+W)X^T$	XGX^T

One possible way to address the SSS problem is by performing PCA projection to reduce the dimension of the feature space and make S_2 to be nonsingular, such as the Laplacianface method in [8] and the DLPP method in [20]. However, there are two shortcomings in this solution:

1) The recognition accuracy depends very much on the dimension of the reduced PCA subspace, and how to determine the optimal dimension of this subspace remains largely an open problem.

2) Some useful information for LPP may be compromised in the intermediate PCA stage. To illustrate this point, we provide the following justification:

In [8], the locality preserving power of a projection w was defined as:

$$f(w) = \frac{w^T XLX^T w}{w^T XD X^T w} \quad (4.2)$$

In general, the smaller the function $f(w)$ is, the better the locality preserving power of the projection w . As the LPP in [8] can be obtained by imposing a normalization constraint $w^T XD X^T w = 1$, Eq. (4.2) is reduced to

$$f(w) = w^T XLX^T w \quad (4.3)$$

Since XLX^T is usually singular [8] in face recognition, which implies that the null space of XLX^T contains useful discriminant information for LPP, and such information could be lost in the PCA projection when $f(w) = 0$. Hence, preprocessing high dimensional features using PCA is arguably not an optimal solution to the SSS problem in LPP.

Nevertheless, the SSS problem has been extensively addressed in classical LDA and many solutions have been proposed. For example, Friedman [56] and Lu *et al.* [59] applied a small perturbation to the within-class scatter; Dai *et al.* [57] proposed a regularized discriminant analysis (RDA) algorithm to regulate the whole space of the within-class

scatter matrix; Jiang *et al.* [58] put forward an eigenfeature regularization and extraction (ERE) approach to decompose the within-class scatter space into three different subspaces and regulate the eigenspectrum using piecewise function. All these algorithms can enhance the recognition performance of LDA to a certain extent. There is however less work on addressing the SSS problem in LPP and its variants. While many manifold learning algorithms have been developed, not much attention has been devoted to the corresponding SSS problem.

Some existing algorithms apply PCA to avoid the SSS problem. But PCA cannot be an optimal solution as we discussed before. Some algorithms ignore the SSS problem [21] and [22], and they may not work as effectively when the sample size is small.

We propose here an effective parametric regularized locality preserving projections (PRLPP) algorithm to overcome or mitigate the SSS problem. Our contribution consists of the following.

- 1) Attaining an efficient subspace which has more locality preserving power than LPP.
- 2) Applying a parametric regularization approach instead of PCA to address the SSS problem for LPP, and extending this regularization technique to other manifold learning algorithms such as DLPP, ODLPP and UDLPP.

The remaining of this chapter is organized as follows. Section 4.2 formally formulates the proposed parametric regularized algorithms. Experimental results on face recognition are shown in Section 4.3. Section 4.4 concludes this chapter.

4.2 Parametric Regularized Subspace Learning

Given a set of face images $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{d \times N}$, where d is the dimension of each face sample and N is the number of samples. Let S be an affinity matrix defined as:

$$S_{ij} = \begin{cases} \exp(-\|x_i - x_j\|_2^2/t), & \text{if } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ 0, & \text{otherwise} \end{cases} \quad (4.4)$$

where $N_k(x)$ denotes the k nearest neighbors of x , and k and t are two empirically pre-specified parameters.

The basis of the proposed PRLPP is obtained by solving the following constrained optimization problem:

$$\begin{aligned} w_{opt} &= \arg \min_w w^T X L X^T w \\ \text{subject to} \quad & w^T H w = 1 \end{aligned} \quad (4.5)$$

where $L = D - S$, D is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$, and H is a full-rank $d \times d$ matrix and an approximation to $X D X^T$, i.e., $H \approx X D X^T$. It follows that the basis of PRLPP comprises the eigenvectors of the following generalized eigenvalue problem:

$$X L X^T w = \lambda H w \quad (4.6)$$

Denote $F = X D X^T$. Our aim is to find a full-rank matrix H that well approximates F . To achieve this goal, we first analyze the spectrum of F using singular value decomposition (SVD), as follows:

$$F = U \Sigma V^T = \sum_{i=1}^d \sigma_i u_i v_i^T = \sum_{i=1}^r \sigma_i u_i v_i^T \quad (4.7)$$

where $U = [u_1, \dots, u_r, u_{r+1}, \dots, u_d]$ and $V = [v_1, \dots, v_r, v_{r+1}, \dots, v_d]$ are two orthogonal matrices, $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_d)$, $\sigma_1 \geq \sigma_2 \geq \dots \sigma_r > \sigma_{r+1} = \dots = \sigma_d = 0$, and $r = \text{rank}(F)$. Here $U = V$ as F is symmetric and positive semi-definite.

To better understand the spectrum of F in face recognition, we examine the F computed from 400 images of 32×32 pixels in the ORL database. Since F is a 1024×1024 matrix and its rank (r) is less than 400, the SSS problem is likely to affect the recognition performance. Let $\Phi^r = [u_1, \dots, u_r]$, $\Phi^n = [u_{r+1}, \dots, u_d]$, and $\Phi^w = [u_1, \dots, u_r, u_{r+1}, \dots, u_d]$ be the range space, null space, and whole space, and σ_i , $i = 1, 2, \dots, d$, be the spectrum of F , respectively. From the plot of spectrum σ_i as shown in Fig. 4.1, it is easy to see that most energy of F concentrates on a few principal components corresponding to the larger eigenvalues in the range space. Those components corresponding to smaller eigenvalues in the range space are considered not reliable due to lack of enough samples. Similar to the analysis for LDA [9], we can decompose F into three subspaces, called the principal space (P), noise space (N) and null space (O), as

$$F = \sum_{i=1}^c \sigma_i u_i u_i^T + \sum_{i=c+1}^r \sigma_i u_i u_i^T + \sum_{i=r+1}^d \sigma_i u_i u_i^T \quad (4.8)$$

Fig. 4.1: Spectrum of F .

where

$$u_i \in \begin{cases} P, & i \leq c \\ N, & c+1 \leq i \leq r \\ O, & r+1 \leq i \leq d \end{cases} \quad (4.9)$$

The value of c can be decided by a pre-specified energy percentage of F , as follows:

$$c = \min\{c \mid \sum_{i=1}^c \sigma_i / \sum_{i=1}^d \sigma_i \geq \zeta\} \quad (4.10)$$

In our experiments, ζ is set to be 98%. In other words, 98% energy of F is retained in the principal space and its components do not require regularization as the discriminant information in this space is considered to be reliable.

Different from the previous work which models the spectrum of the range space of F as a function in the form of $1/f$ [58,60], we extend it to a more general form of parametric regularization: $1/f^a$ ($a > 0$), where f is the index number of the eigenfeature spectrum. Furthermore, we introduce another parameter b to model the null space rather than a deterministic constant in [58] and the extrapolated eigenvalues in [60]. Our parametric modeling stems from the two following considerations:

1) The function $1/f$ may not precisely model the spectrum in terms of the modeling error and recognition accuracy. In general, the modeled spectrum should be as similar

Fig. 4.2: The real spectrum and the modeled spectrum with different value of regularization parameter a .

to the real spectrum as possible so that the modeling error can be minimized. Fig. 4.2 shows the spectrum computed from the ORL database and the modeled spectrum with different values of parameter a , along with the corresponding mean square error (MSE) defined as

$$MSE = \frac{1}{r} \sum_{i=1}^r \left\| \frac{\alpha}{(i + \beta)^a} - \sigma_i \right\|_2 \quad (4.11)$$

We can see that the optimal parameter for modeling the spectrum is equal to 0.8 rather than 1.0 as suggested in [58, 60].

2) Since we may not have any prior information on the null space, modeling it in a parametric manner is more suitable than using a deterministic constant, we can estimate the optimal regularization parameter according to the recognition accuracy achieved during the training phase. Fig. 4.3 shows the recognition accuracy versus different values of parameter b ranging from 0 to 400. We can observe that an appropriate b is very important to the recognition accuracy. Hence, applying a parametric regularization approach

Fig. 4.3: Recognition accuracy versus different value of parameter b .

can likely lead to a better performance in recognition.

By formulating F into the proposed parametric form, we have

$$F' = \sum_{i=1}^c \sigma_i u_i u_i^T + \sum_{i=c+1}^r \frac{\alpha}{(i + \beta)^a} u_i u_i^T + \sum_{i=r+1}^d b u_i u_i^T \quad (4.12)$$

Assume that $\sigma_1 = \alpha/(1 + \beta)^a$ and $\sigma_c = \alpha/(c + \beta)^a$. It follows that

$$\alpha = \frac{\sigma_1 \sigma_c (c - 1)^a}{(\sigma_1^{1/a} - \sigma_c^{1/a})^a} \quad (4.13)$$

$$\beta = \frac{c \sigma_c^{1/a} - \sigma_1^{1/a}}{\sigma_1^{1/a} - \sigma_c^{1/a}} \quad (4.14)$$

We now discuss how to determine the optimal parameters a and b in practice. For a given dataset, we can randomly select 3 samples of each subject, and use 2 of them to construct the training set and the remaining one as the testing set. We determine the optimal a through solving the following optimization equation:

$$a_{opt} = \arg \min_a \sum_{i=1}^r \left\| \frac{\alpha}{(i + \beta)^a} - \sigma_i \right\|_2 \quad (4.15)$$

As α and β are both functions of a , it is nontrivial to obtain a closed-form solution to (4.15). Hence, we solve for a in an iterative manner, as described below:

Step 1: Set the iteration step size $\eta = 0.05$ and initialize $a_0 = 1.0$.

Step 2: Compute $err_0 = \sum_{i=1}^r \|\frac{\alpha}{(i+\beta)^{a_0}} - \sigma_i\|_2$, the initial error between the real spectrum and the modeled spectrum.

Step 3: Set the convergence precision $\epsilon = err_0/1000$.

Step 4: Set $k = 1$, $a_k = a_{k-1} + \eta$, and compute $err_k = \sum_{i=1}^r \|\frac{\alpha}{(i+\beta)^{a_k}} - \sigma_i\|_2$.

Step 5: Repeat until $|err_k - err_{k-1}| < \epsilon$

If $err_k < err_{k-1}$

Set $a_{k+1} = a_k + \eta$;

Otherwise ($err_k \geq err_{k-1}$)

Set $a_{k+1} = a_k - \eta$

Set $k = k + 1$ and compute $err_k = \sum_{i=1}^r \|\frac{\alpha}{(i+\beta)^{a_k}} - \sigma_i\|_2$.

Step 6: Output $a = a_k$.

After determining the optimal value for a , we apply a cross validation procedure as illustrated in Fig. 4.3 to search for the optimal b corresponding to the highest recognition accuracy.

Let $H = F'$, and we can obtain the basis functions of PRLPP by solving the generalized eigenvalue equation as defined in (4.6). Let the column vector $w_1, w_2, \dots, w_l$ be the solutions of (4.6) and ordered according to eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_l$. The embedding of PRLPP can be obtained as follows:

$$y_i = W^T x_i \quad (4.16)$$

where $W = [w_1, w_2, \dots, w_l]$, a $d \times l$ matrix.

We can extend this regularization algorithm to regulate the S_2 of DLPP, ODLPP and UDLPP, and formulate the regularized PRDLPP, PRODLPP and PRUDLPP in the same way.

Fig. 4.4: Sample face images from the ORL database.

4.3 Experimental Results

We have evaluated the effectiveness of our proposed PRLPP and its extension to other manifold learning algorithms on face recognition. Two publicly available face databases, namely ORL and FERET [107, 108], were used in our evaluation. The ORL database contains 400 images from 40 persons, and each person has 10 different images acquired at different times. The images were taken with a tolerance for some tilting and rotation of the face up to 20° and some variation in the scaling up to 10%. All images are gray and normalized to a resolution of 92×112 pixels. For the sake of simplicity and computational efficiency, we manually aligned and cropped each face image according to the eyes' positions and resized it into 32×32 pixels.

The FERET database comprises 13539 face images corresponding to 1565 subjects. Similar to the previous work in [61], we selected a subset of 1400 face images of 200 subjects who have the same set of pose from the FERET face database. For each subject, there were 7 different images whose names are marked with two-character strings: "ba", "bj", "bk", "be", "bf", "bd" and "bg" to denote different face poses. Different from the ORL database, where each face image was well-aligned, the face images of the FERET database in the experiments were not well aligned. The facial part of each image was detected by Adaboost face detector and then roughly cropped and resized into 32×32 pixels without using the precise eyes' positions. The aim here is to examine the recognition performance of the proposed method on misaligned face images. Sample images of one subject from the ORL and FERET database are shown in Figs. 4.4 and 4.5, respectively.

Fig. 4.5: Sample face images from the FERET database.

4.3.1 Locality Preserving Power

Both LPP and PRLPP try to preserve the local geometrical structure. They find the basis vector w by minimizing the locality preserving function [11]

$$f(w) = \frac{w^T S_1 w}{w^T S_2 w} \quad (4.17)$$

where $S_1 = XLX^T$, $S_2 = XDX^T$ for LPP and $S_2 = H$ for PRLPP, $f(w)$ reflects the locality preserving power of the projection w . As $f(w)$ is to be minimized, the smaller the function $f(w)$ is, the better the locality preserving power of the projection w .

Based on the Rayleigh quotient format of the eigenvalue problem in (4.1) and (4.6), the value of $f(w)$ in (4.17) is exactly the eigenvalue of (4.1) and (4.6) corresponding to eigenvector w . Therefore, the eigenvalues of LPP and PRLPP reflect the locality their locality preserving power and smaller eigenvalues have more locality preserving power.

Fig. 4.6 shows the eigenvalues of LPP and PRLPP using the ORL database. As can be seen from the figure, the eigenvalues of PRLPP are consistently smaller than those of LPP, which indicates that PRLPP has more locality preserving power than LPP. The superiority of PRLPP is due to the fact that the whole space of F have been utilized for feature representation in PRLPP, while LPP discards the small and zero eigenvalues.

4.3.2 Face Recognition

We conducted experiments of face recognition to evaluate the efficacy of PRLPP and compared it with PCA [1], LDA [2] and LPP [8] in terms of recognition accuracy. We also applied some existing regularization methods developed for LDA to regulate LPP and compare their performance with our proposed PRLPP method. Lastly, we extended our proposed parametric regularization technique to other manifold learning algorithms including DLPP, ODLPP and UDLPP. We applied the nearest neighbor rule in the Euclidean space to perform recognition. Note that the recognition accuracy reported

Fig. 4.6: Eigenvalues of LPP and PRLPP. The eigenvalues of PRLPP are consistently smaller than those of LPP, which indicates that PRLPP has more locality preserving power than LPP.

here is the rank-1 recognition rate of each experiment. It is defined as N_c/N_t , where N_c is the number of correctly recognized testing samples and N_t is the number of total testing images. In our experiments, to reduce the variation in the recognition results, we computed the average recognition accuracy (ARA) of 20 runs.

1) *Performance of PRLPP*: We randomly selected 5 and 3 images of each subject in the ORL and FERET databases, respectively, to construct the training set, and used the remaining as the testing set. In our experiments, the value of t was empirically set to be 800 for the ORL database and 1000 for the FERET database, respectively, and k is set to be $l - 1$, where l is the number of samples of each subject used for training.

The optimal parameters a and b were determined during the training phase and their values were found to be (0.7, 130) and (0.8, 80), respectively.

We first performed regularization on 1) the principal space, 2) the range space (principal space + noise space), and 3) the whole space (principal space + noise space + null space) of the LPP space. Obviously, if only the principal space is used, our PRLPP method degenerates into the conventional LPP algorithm. The recognition accuracies corresponding to these three different regularizations are shown in Fig. 4.7. It is easy to see that regularization on the whole space always attains the highest recognition accu-

Fig. 4.7: Performance comparison on three different spaces.

Table 4.2: Comparison on recognition accuracy using the ORL and FERET databases.

Method	ORL		FERET	
	Accu. (%)	Dim	Accu. (%)	Dim
PCA	92.5	50	34.6	120
LDA	93.5	39	35.8	100
LPP	94.0	40	65.0	160
PRLPP	96.5	40	69.9	160

racy, while regularization on the range space performs better than that of the principal space.

Further, we compared the recognition accuracy of PRLPP with those of PCA, LDA and LPP, and Table 4.2 tabulates the performance on the two databases. Fig. 4.8 shows the ARAs at different feature dimensions. It is evidenced from the results that our proposed PRLPP algorithm consistently and notably outperforms the other methods under comparison.

We then conducted experiments with different number of training samples for each subject to examine under what condition the proposed PRLPP method can expect more gain. We randomly selected l samples from each subject for training and used the remaining for testing. Fig. 4.9 shows the ARA obtained on the two databases. The results show that when the number of training samples per subject is small, the proposed PRLPP can obtain more gain as compared with LPP.

(a) ORL database.

(b) FERET database.

Fig. 4.8: ARA comparison of PRLPP with PCA, LDA, and LPP.

2) *Comparison with other regularization methods:* In this experiment, we compared the performance of the proposed method with existing regularization methods developed for LDA. We directly applied them to perform regularization for LPP on the ORL and FERET databases, in which 5 and 3 samples of each subject were respectively and randomly selected for training and the remaining were used for testing. The recognition results are listed in Table 4.3, which shows that our proposed parametric regularization

(a) ORL database.

(b) FERET database.

Fig. 4.9: ARA versus different number of training samples of each subject.

Table 4.3: Comparison with existing regularization methods.

Method	ORL		FERET	
	Accu. (%)	Dim	Accu. (%)	Dim
Regularized LDA [56]	94.0	60	55.5	180
Dai’s method [57]	95.0	40	67.3	160
PRLPP	96.5	40	69.9	160

Table 4.4: Performance comparison of three manifold learning algorithms and our proposed regularized extensions.

Method	DLPP	ODLPP	UDLPP	PRDLPP	PRODLPP	PRUDLPP
Accu. (%)	88.6	89.2	89.5	89.7	90.0	90.7

method notably performs better than the existing methods.

3) *Extensions to other manifold learning methods:* In this experiment, we extended the proposed regularization algorithm to three other manifold learning algorithms—DLPP, ODLPP and UDLPP—and formulated the PRDLPP, PRODLPP and PRUDLPP in the same way. We compared the regularized algorithms with the original algorithms on the ORL database. We randomly selected l samples of each subject for training and the remaining for testing. Table 4.4 lists the recognition accuracy achieved when l is equal to 3, and Fig. 4.10 shows the performance with different number of training samples per subject.

4) *Computational complexity:* We examined the computational cost of the proposed PRLPP method and compared it with those of other subspace learning methods including PCA, LDA and LPP. Our hardware configuration is 2.66-GHz CPU and 2G RAM. Table 4.5 shows the CPU time spent on the training and the testing (recognition) phases by these subspace learning methods using Matlab software and the ORL database. Note that 5 samples of each subject on this database were selected for training and the dimension of all the reduced subspaces was selected to be 39 in this experiment.

Table 4.5: CPU time (in second) used by different methods on the ORL database.

Method	PCA	LDA	LPP	PRLPP
Training	0.234	0.234	0.344	11.350
Recognition	0.001	0.001	0.001	0.001

Fig. 4.10: Performance comparison with different number of training samples per subject.

The computational complexity of the proposed algorithm for training is generally larger than other subspace learning methods as it extracts features in a whole space rather than a reduced PCA subspace such as LDA and LPP. In practical applications, however, training is usually an offline process and only recognition needs to be performed in real time. Thus, the recognition time is usually more of our concern than the training time. As shown in Table 4.5, the recognition time of the proposed algorithm is the same as other approaches for the same number of features. Moreover, the proposed algorithm can be more computationally efficient in recognition than other approaches including PCA, LDA and LPP, for the reason that it requires a smaller number of features than other approaches to obtain the same recognition accuracy (see Fig. 4.8).

4.3.3 Discussion

The following observations can be made from the experimental results listed in Tables 4.2-4.4 and Figs. 4.6-4.10:

1. PRLPP consistently obtains higher recognition accuracy than PCA, LDA and LPP. Compared with PCA and LDA, PRLPP can better preserve the nonlinear manifold structure as LPP, which is more suitable for large-scale face recognition, and it is evidenced from the results in Table 4.2 on the FERET database. The gain of PRLPP over the best recognition accuracy of PCA and LDA are 3.0% and 34.1% on the ORL and

FERET databases, respectively. Compared with LPP, PRLPP doesn't need to perform PCA first and the information in the whole space is fully exploited to achieve better recognition accuracy.

2. PRLPP achieves more gain than LPP, especially when the number of images of each subject in the training set is small. For example, when only 2 images of each subject are used for training, the gains of PRLPP over LPP are 3.9% and 6.0% on the ORL and FERET databases, respectively. The gains decrease to 2.5% and 0.6%, respectively, when 5 images of each subject are used for training. The reason is that when the number of samples of each subject is small, more information is compromised in the PCA preprocessing step and the small sample size problem becomes more prominent.

3. PRLPP achieves better recognition accuracy than other existing regularization algorithms developed for LDA. Compared with the RDA methods in [56] and [57], PRLPP performs regularization in the noise space and null space, rather than those applied in the whole space of the within-class scatter matrix, and can thus better preserve the principal space. Compared with the ERE method in [58], our proposed PRLPP method performs regularization in a parametric manner, employing the optimal parameters obtained during the training phase rather than using deterministic constants. Empirical results also show that the optimal parameters are usually not the constants as used in the ERE method.

4. The proposed parametric regularization can be extended to other manifold learning algorithms such as DLPP, UDLPP and ODLPP, resulting in improved performance. For instance, when 3 samples of each subject are used for training on the ORL database, the regularized methods—PRDLPP, PRUDLPP and PRODLPP—achieve 1.1%, 0.8% and 1.2% gains, respectively, compared with the original methods without regularization. Moreover, when the proposed regularization method is applied to other LPP-based methods, such as DLPP, ODLPP and UDLPP, we can achieve better performance than those applied to conventional LPP. For example, when 3 samples of each subject are used for training on the ORL database, the regularized methods—PRDLPP, PRUDLPP and PRODLPP—achieve 0.7%, 1.0% and 1.7% gains over PRLPP, respectively (see Fig. 4.9(a) and Fig. 4.10). The reason is that other LPP-based methods such as DLPP,

ODLPP and UDLPP are originally supervised. Hence, both locality geometrical information and discriminant information have been simultaneously exploited in these methods and the subspaces learnt by these methods can have more discriminant power than that of LPP. Accordingly, performing regularization on these LPP-based methods also outperforms applying that on the conventional LPP method.

4.4 Chapter Summary

We have proposed in this chapter a parametric regularized locality preserving projections (PRLPP) method for recognition. The proposed method is designed to regulate LPP features corresponding to small and zero eigenvalues, i.e., the noise space and null space, and exploit the discriminant information from the whole space, resulting in more locality preserving power and better recognition performance than LPP. Moreover, the proposed parametric regularization method can also be extended to other manifold learning algorithms and similar performance improvements have been obtained. Furthermore, our proposed method appears to be the first algorithm that formally addresses the SSS problem in manifold-based subspace learning using parametric regularization analysis. Partial work presented in this chapter was published in [163].

Chapter 5

Cost-Sensitive Subspace Learning

5.1 Introduction

Existing subspace learning-based face recognition algorithms aim to seek low recognition errors and implicitly assume same loss of all misclassifications. In many real-world face recognition applications, however, this assumption may not hold as different misclassifications could lead to different losses. For example, it may cause inconvenience to a gallery person who is mis-recognized as an impostor and not allowed to enter the room by a face recognition-based door-locker, but may result in a serious loss or damage if an impostor is mis-recognized as a gallery person and allowed to enter the room. Another example is the face sketch recognition-based criminal search system. It could cause some trouble to an innocent citizen who is mis-recognized as a criminal suspect by the criminal search system, but may result in a large insecurity if a criminal suspect is mis-recognized as innocent citizen and allowed to go free. From these two typical application examples, we can conclude that face recognition is a cost-sensitive pattern classification problem and it is not very reasonable, at least not optimal, to only seek the minimal recognition error in many practical face recognition applications.

Cost-sensitive learning is one important and hot research topic in the data mining and machine learning community, and has been studied for years [62–68]. In such settings, cost information is introduced to measure the importance of different samples in different classes, and different costs reflect different amounts of losses. The aim of cost-sensitive learning is to minimize the total cost rather than the total error. Following this basic idea, several cost-sensitive learning methods have been proposed in the literature,

such as cost-sensitive boosting [63, 64], cost-sensitive support vector machine [62], cost-sensitive semi-supervised learning [62], and cost-sensitive neural networks [66]. While these cost-sensitive learning methods can attain better performance than conventional corresponding uniform-cost methods, and have been successfully applied to many machine learning and data mining problems, little progress has been made to address the cost-sensitive issue in face recognition.

To consider the difference between a false rejection and false acceptance, many previous face recognition systems applied the well-known receiver operating characteristic (ROC) curve to choose an appropriate threshold for classification [69]. These methods can be seen as implicitly utilizing the cost information of different samples because the threshold depends on the cost selected. To the best of our knowledge, Reference [68] could be the first work that formally addressed the different losses in a face recognition system by formulating face recognition as a multi-class cost-sensitive classification problem. While improved recognition performance (smaller recognition loss) has been obtained, their method uses only the cost-sensitive learning technique in the classification phase and not in the feature extraction (subspace learning) phase. In other words, the samples with different costs have not been utilized in the feature extraction phase, hence useful cost-sensitive information may be ignored in this phase and result in suboptimal solution.

In this paper, we propose a new cost-sensitive subspace learning approach for face recognition. Different from existing cost-sensitive learning techniques, our aim now is to derive cost-sensitive features for face representation and recognition. Our approach incorporates a cost matrix, which specifies the different costs associated with misclassifications of different subjects, into three popular subspace learning algorithms and devise the corresponding cost-sensitive methods, namely, cost-sensitive principal component analysis (CSPCA), cost-sensitive linear discriminant analysis (CSLDA), and cost-sensitive locality preserving projections (CSLPP), in such a way that samples with higher costs are emphasized and those with lower costs are deemphasized in both the learning and testing phases. Experimental results are presented to demonstrate the efficacy of the proposed approach.

The remaining of this chapter is organized as follows. Section 5.2 details our proposed cost-sensitive subspace learning approach for face feature extraction. The experimental

results and discussions are presented in Section 5.3. We conclude the chapter and suggest the future work in Section 5.4.

5.2 Cost-Sensitive Subspace Learning

5.2.1 Motivation

In a practical face recognition system, there are usually three types of misclassifications on recognizing a test face sample:

- 1) **False rejection**, mis-recognizing a gallery person as an impostor;
- 2) **False acceptance**, mis-recognizing an impostor as a gallery person;
- 3) **False identification**, mis-recognizing a gallery person as another gallery person.

In most existing face recognition systems, these errors are treated equally and assumed to cause the same loss. It is easy to see that the loss due to a false acceptance to an access controlled system could be far greater than a false rejection. Recognizing a gallery person as another gallery person or an impostor will likely result in a different loss. Motivated by this reason, we propose to explicitly incorporate the possible different losses caused by different classification errors into the feature extraction (subspace learning) phase of a face recognition system.

Let $X = [x_1, x_2, \dots, x_{N_1}, x_{N_1+1}, \dots, x_{N_1+N_2}]$ be a face image set, where $x_i \in R^d$, $1 \leq i \leq N_1 + N_2$, N_1 and N_2 are the numbers of the gallery and impostor images, respectively, and d is the feature dimension of each face sample. The class label of x_i is assumed to be $l_i \in \{G_1, G_2, \dots, G_M, I\}$, where G_j is the gallery class, $j = 1, 2, \dots, M$, and I is the impostor class, a general meta-class. Let C_{t_1, t_2} be the cost incurred by recognizing a face sample of the t_1 th class as the t_2 th class, where $1 \leq t_1, t_2 \leq M + 1$, the first M classes are the gallery persons and the $M + 1$ th class are the impostor subjects. For a practical face recognition system, we can construct a cost matrix C as shown in Table. 5.1, where C_{t_1, t_2} denotes the cost of recognizing a face sample of the t_1 th class as the t_2 th class. The diagonal elements in the cost matrix are zero because there is no loss for a correct recognition.

The cost C_{t_1, t_2} of different types of misclassifications are set in the non-diagonal elements of the cost matrix C , described as follows:

Table 5.1: Cost matrix of a practical face recognition system.

	G_1	$\cdots$	G_M	I
G_1	0	$\cdots$	$C_{1,M}$	$C_{1,M+1}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
G_M	$C_{M,1}$	$\cdots$	0	$C_{M,M+1}$
I	$C_{M+1,1}$	$\cdots$	$C_{M+1,M}$	0

- 1) False rejection: the $M+1$ th column of C except the last element, where $t_2 = M+1$;
- 2) False acceptance: the $M+1$ th row of C except the last element, where $t_1 = M+1$;
- 3) False identification: the upper-left $M \times M$ block in C except the diagonal elements, where $1 \leq t_1, t_2 \leq M$.

While the proposed framework allows all possible forms of loss for different applications, we assume that the cost matrix is given by the users and we will focus on how to extract cost-sensitive features to improve the performance of a practical face recognition system.

5.2.2 Proposed Methods

While a large number of subspace learning-based face recognition algorithms have been proposed in the literature, PCA, LDA and LPP are unarguably the most representative and popular three methods, and they are also called the well-known Eigenfaces, Fisherfaces and Laplacianfaces in the face recognition community. Hence, we incorporate the cost matrix into these three methods, and formulate the corresponding CSPCA, CSLDA, and CSLPP methods to extract cost-sensitive features for face representation and recognition. Besides these three methods, the basic idea of our cost-sensitive subspace learning approach can also be easily applied to other subspace learning methods to derive the corresponding cost-sensitive variations similar to that used in this paper. However, it is out of the key scope of this chapter.

Let $N = N_1 + N_2$ and redesignate X as $X = [x_1, x_2, \cdots, x_N]$. Let $W = [w_1, w_2, \cdots, w_k]$ be the cost-sensitive subspace to be sought, and W map x_i into a low dimensional feature y_i , where $y_i = W^T x_i$, $y_i \in R^k$, $k \ll d$. The essential differences among different cost-sensitive subspace learning methods lie in their differences in defining and finding the projection matrix W .

1) *CSPCA*: We define the objective of CSPCA as follows:

$$\max_w J_T(w) = \frac{1}{2} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \text{cost}(y_i, y_j) \|y_i - y_j\|^2 \quad (5.1)$$

where $\text{cost}(y_i, y_j)$ indicates the cost of mis-recognizing y_i as y_j .

Let x_i and x_j belong to the t_1 th and t_2 th classes in the training set, where $1 \leq t_1, t_2 \leq M + 1$. It follows that $\text{cost}(y_i, y_j) = \text{cost}(x_i, x_j) = C_{t_1, t_2}$ and it can be easily obtained from the cost matrix C . After some algebraic manipulations, we can simplify $J_T(w)$ to the following form

$$\begin{aligned} J_T(w) &= \frac{1}{2} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \text{cost}(x_i, x_j) \|w^T x_i - w^T x_j\|^2 \\ &= w^T \left[\frac{1}{2} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \text{cost}(x_i, x_j) (x_i - x_j)(x_i - x_j)^T \right] w \end{aligned} \quad (5.2)$$

Let $S_T(w) = \frac{1}{2} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \text{cost}(x_i, x_j) (x_i - x_j)(x_i - x_j)^T$, The projections $\{w_1, w_2, \dots, w_k\}$ comprise an orthogonal set of vectors which are eigenvectors of $S_T(w)$ associated with the k ($k < d$) largest eigenvalues.

2) *CSLDA*: To characterize the cost of each class, we first define an importance function $f(k)$ to depict the importance of the samples in the k th class, where $1 \leq k \leq M + 1$. Generally, the larger the cost of the k th class, the more importance the class is. Hence, $f(k)$ can be defined as

$$f(k) = \sum_{t_2=1}^{M+1} C_{k, t_2} \phi(k - t_2) \quad (5.3)$$

where $\delta(n)$ is a discrete unit impulse function, and defined as

$$\phi(n) = \begin{cases} 0, & n = 0 \\ 1, & n \neq 0 \end{cases} \quad (5.4)$$

We then define the between-class scatter S_B and the within-class scatter S_W of CSLDA as follows:

$$S_B = \sum_{t_1=1}^{M+1} \sum_{t_2=1}^{M+1} \text{cost}(t_1, t_2) (m_{t_1} - m_{t_2})(m_{t_1} - m_{t_2})^T \quad (5.5)$$

$$S_W = \sum_{k=1}^{M+1} \sum_{l(x_i)=k} f(k)(x_i - m_k)(x_i - m_k)^T \quad (5.6)$$

where m_{t_1} and m_{t_2} are the means of the t_1 th and t_2 th classes, respectively, l_{x_i} is the label of x_i , $cost(t_1, t_2)$ is the cost of mis-recognizing the face images of the t_1 th class into the t_2 th class, and it can be directly obtained from the cost matrix as shown in Table 5.1.

With S_B and S_W , we can obtain the bases of CSLDA by solving the following generalized eigenvalue equation:

$$S_B w = \lambda S_W w \quad (5.7)$$

Let $\{w_1, w_2, \dots, w_k\}$ be the eigenvectors corresponding to the k largest eigenvalues $\{\lambda_i | i = 1, 2, \dots, k\}$ ordered such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$. Then $W = [w_1, w_2, \dots, w_k]$ is the subspace projection of CSLDA. In fact, S_W is usually singular in face recognition, which stems from the fact that the number of training images is usually much smaller than the dimension of each image, a deficiency that is generally known as small sample size (SSS) problem. One possible way to address this SSS problem is to perform PCA projection to reduce the dimension of the original face feature space. It is to be noted that it is not suitable for CSLDA because some useful cost-sensitive information for CSLDA may be compromised in the intermediate PCA stage, as discussed in the last chapter. Hence, we alternatively applied the regularization technique proposed in [56] to add a small perturbation μI to the within-class scatter of CSLDA. In our experiments, μ is empirically set to be 0.001.

3) *CSLPP*: We define the objective of CSLPP as follows:

$$\min \sum_{i=1}^N \sum_{j=1}^N cost(y_i, y_j) \|y_i - y_j\|^2 S_{ij} \quad (5.8)$$

where $cost(y_i, y_j)$ indicates the cost of mis-recognizing y_i as y_j and the definition of S is an affinity matrix defined as:

$$S_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2/t), & \text{if } x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i) \\ 0, & \text{otherwise} \end{cases} \quad (5.9)$$

where $N_k(x)$ denotes the k nearest neighbors of x , and k and t are two empirically pre-specified parameters.

Let x_i and x_j belong to the t_1 th and t_2 th classes in the training set, $1 \leq t_1, t_2 \leq M+1$. It follows that $\text{cost}(y_i, y_j) = \text{cost}(x_i, x_j) = C_{t_1, t_2}$ and it can be also easily obtained from the cost matrix C . With some algebraic manipulation, the objective function can be simplified to the following form

$$\begin{aligned}
& \sum_{i=1}^N \sum_{j=1}^N \text{cost}(y_i, y_j) \|w^T x_i - w^T x_j\|^2 S_{ij} \\
&= \sum_{i=1}^N \sum_{j=1}^N \text{cost}(x_i, x_j) w^T (x_i - x_j) (x_i - x_j)^T w S_{ij} \\
&= 2 \sum_{i=1}^N \text{cost}(x_i, x_i) w^T x_i S_{ij} x_i w - 2 \sum_{i=1}^N \sum_{j=1}^N \text{cost}(x_i, x_j) w^T x_i S_{ij} x_j^T w \\
&= -2w^T A w
\end{aligned} \tag{5.10}$$

where $A = \sum_{i=1}^N \sum_{j=1}^N \text{cost}(x_i, x_j) x_i S_{ij} x_j^T$. To make the minimization problem with respect to w well-posed, we impose a constraint $w^T w = 1$ and formulate CSLPP as the following constrained optimization problem:

$$\begin{aligned}
& \max_w \quad w^T A w \\
& \text{s.t.} \quad w^T w = 1.
\end{aligned} \tag{5.11}$$

We can then obtain the bases of CSLPP by solving the following eigenvalue equation

$$A w = \lambda w \tag{5.12}$$

Let $\{w_1, w_2, \dots, w_k\}$ be the eigenvectors corresponding to the k largest eigenvalues $\{\lambda_i | i = 1, 2, \dots, k\}$ ordered such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$. Then $W = [w_1, w_2, \dots, w_k]$ is the subspace projection of CSLPP.

5.2.3 Discussion

In this subsection, we discuss the relationship between our proposed cost-sensitive subspace learning methods and two related works: the existing subspace learning methods [1, 2, 8] and cost-sensitive classifiers [68].

Relationship with the existing subspace learning methods: Existing subspace learning-based face recognition methods aim to seek a low recognition error rate and implicitly

assume that the losses of all misclassifications are the same. Hence, they can be considered as uniform-cost methods. As a result, our proposed cost-sensitive methods perform recognition to attain low losses rather than low recognition errors, which could be more suitable for many real-world face recognition applications. It is obvious to observe that existing uniform-cost subspace learning methods such as PCA, LDA and LPP can be seen as special cases of our proposed CSPCA, CSLDA and CSLPP when we simply set $cost(x_i, x_j) = 1$ for arbitrary two samples in the training set, where $1 \leq i, j \leq N$.

Recently, Zhang *et al.* [68] proposed two cost-sensitive classification methods, called multi-class cost-sensitive kernel logistic regression (mcMLR) and multi-class cost-sensitive k -nearest neighbor (mckNN) for cost-sensitive face recognition. While improved recognition performance (smaller recognition loss) can be obtained, the methods apply the cost-sensitive learning techniques only in the classification phase and not in the learning (feature extraction) phase since they make use of the conventional PCA, LDA and LPP features for recognition. In other words, the samples with different costs have not been utilized in the feature extraction phase, and some useful cost-sensitive information may not be exploited in this phase. Our proposed approach has explicitly exploited the cost-sensitive information of the samples from different classes in the training phase rather than in the classification phase. Moreover, our empirical results also show that the performance of the proposed cost-sensitive subspace learning methods can be further improved when using the cost-sensitive classifiers for recognition. More details are provided in next section.

5.3 Experimental Results

We present in this section a series of experiments with a synthetic dataset and real face databases to evaluate the efficacy of the proposed cost-sensitive subspace learning approach. The first experiment on the synthetic dataset show how the cost matrix affects the feature (eigenvector) extracted by existing subspace learning algorithms in a cost-sensitive recognition problem. Experiments on real face data show the efficacy of the proposed approach for face representation and recognition.

Table 5.2: Cost matrix used in the toy example. Class 1, 2, and 3 denote the cross, circle and square class, respectively.

	Class 1	Class 2	Class 3
Class 1	0	100	100
Class 2	10	0	1
Class 3	1	10	0

5.3.1 Toy Example

To demonstrate the effectiveness of exploiting different costs associated with misclassifications of the samples in different classes, we constructed a synthetic dataset containing three different classes according to the following distribution:

$$\begin{cases} x_k^i = x_0^i + \delta_1(k) \\ y_k^i = y_0^i + \delta_2(k) \end{cases} \quad (5.13)$$

where $i = 1, 2, 3$ denotes the class index, $x_0^1 = -2$, $x_0^2 = 1.5$, $x_0^3 = 1.5$, $y_0^1 = 0$, $y_0^2 = -2$, $y_0^3 = 3$, $k = 1, 2, \dots, 25$, $\delta_1(k)$ and $\delta_2(k) \sim N(0, 1)$. Fig. 5.1 shows the data of the three classes generated by the above distribution, which are denoted by crosses, circles and squares, respectively. Assume the cross class denotes the imposter class, and the circle and square classes denote two gallery subjects in a face recognition-based access control system. Hence, the cost of mis-recognizing the cross samples as the other two classes is generally higher than that of mis-recognizing the samples in the circle and square classes as the cross class. Based on this reason, we assigned different costs of these samples as shown in Table. 5.2. Fig. 5.1 shows the eigenvectors sought by LDA and CSLDA on this dataset, We can observe from this figure that the projection direction sought by CSLDA is different from that of LDA. Specifically, the projection direction of LDA can better separate these three classes than that of CSLDA because the circle and the square classes were heavily overlapped on the projected space of CSLDA. However, the cross class was better separated from the other two classes by CSLDA, and the samples in this class were less misclassified. Since the cost of misclassifying the samples in the cross class is much larger than that of the other classes, CSLDA can attain lower cost than LDA.

To better quantify the performance of LDA and CSLDA, we calculated the total cost and the total error rate (TER) obtained by using the nearest neighbor rule and tabulated

Fig. 5.1: The projection direction sought by LDA and CSLDA, respectively.

Table 5.3: Comparison of total cost and total error rate (TER) on synthetic data set.

Method	cost	TER (%)
LDA	804	16.0
CSLDA	646	40.0

in Table. 5.3. We can see from this table that CSLDA can obtain less cost than LDA while LDA can attain less TER than CSLDA, which indicates that the cost-sensitive method emphasizes more the samples with larger cost in the learning procedure to make the samples of this class be less misclassified, such that the total cost can be reduced to enhance the security of the system.

5.3.2 Face Representation and Recognition

In this subsection, we evaluate the effectiveness of our proposed cost-sensitive subspace learning methods for face representation and recognition.

Data Preparation: We made use of the AR [109] and FERET [107, 108] face databases in our experiments. The AR database contains over 4000 face images of 126 subjects (70 men and 56 women), including frontal facial images with different facial expressions, lighting conditions and occlusions. In our experiments, we selected a subset of 1540 face images of 110 subjects with 14 face images of different expressions and lighting condi-

Fig. 5.2: Samples of the cropped and aligned face image from the (a) AR database and (b) FERET database.

tions per subject. The FERET database contains 1199 subjects with a total of over 14000 images. These images contain different lighting variations, facial expressions, and pose angles. In our experiments, we only selected images of frontal view with different expressions and illuminations. The facial part of each image was manually cropped into 32×32 pixels and aligned according to the eyes' positions. Some samples images of one gallery subject from the AR database and seven impostor subjects from the FERET database are shown in Fig. 5.2. In our experiments, each color facial image was transformed into a gray-scale one by using the function *rgb2gray* in Matlab R2007a.

While our proposed cost-sensitive face recognition framework allows all possible forms of loss, for convenience of our discussion, we assume that accepting any impostor will cause the same loss, and mis-recognizing a gallery person as another gallery person or an impostor will cause different amounts of loss. Without any a priori information or preference, this assumption is reasonable for most practical face recognition systems. Let C_{GI} , C_{IG} and C_{GG} be the costs incurred by a false rejection, false acceptance and false identification, respectively. Without loss of generality, we assume that $C_{GI} = \frac{C_{GI}}{C_{GG}}$, $C_{IG} = \frac{C_{IG}}{C_{GG}}$ and $C_{GG} = 1$ as this will not change the final results.

Face Representation: The parameters M , N_G , N_I , C_{IG} , C_{GI} and C_{GG} are specified in Table. 5.4, where C_{ij} denotes the cost of mis-recognizing a face image of the i th class as the j th class. The training set includes N_G images with M gallery subjects and N_I images as impostors randomly selected from the whole database. We conducted face representation experiments on the AR database to visualize the features extracted by the PCA [1], LDA [2] and LPP [3] methods, as well as their cost-sensitive variants. Fig. 5.3 shows the first 10 bases of PCA, LDA, LPP, CSPCA, CSLDA and CSLPP, respectively.

Table 5.4: Experimental settings for face representation.

M	N_G	N_I	$C_{IG} : C_{GI} : C_{GG}$
30	7	350	20:2:1

Fig. 5.3: First 10 basis images of (a) PCA, (b) LDA, (c) LPP, (d) CSPCA, (e) CSLDA, and (f) CSLPP methods.

It is interesting to note from this figure that the bases of the LDA and CSLDA are more similar to each other than those of PCA and CSPCA, and of LPP and CSLPP. The reason is that PCA and LPP are unsupervised and LDA is supervised, and our proposed cost-sensitive subspace learning is supervised as the label information of each sample has been unitized in the training phase. Hence the bases of LDA and CSLDA are more similar to each other than those of PCA and CSPCA, or of LPP and CSLPP.

Face Recognition: In this subsection, we conducted face recognition experiments on the AR and FERET databases to investigate the cost-sensitive issue in practical face recognition applications and show the efficacy of our proposed approach. We present two different face recognition applications: access control and criminal search.

Application 1—Access Control: For a face recognition-based access control system, it is obvious that $C_{IG} > C_{GI} > C_{GG}$. We set the parameters as shown in Table. 5.5, and conducted face recognition on these two databases, respectively. For each database,

Table 5.5: Parameter settings for a face recognition-based access control system.

Database	M	N_G	N_I	$C_{IG} : C_{GI} : C_{GG}$
AR	30	7	350	20:2:1
FERET	70	3	350	20:2:1

Table 5.6: Comparison of total cost (cost), total error rate (TER), false rejection rate (FRR) and false acceptance rate (FAR) on the AR and FERET databases for the access control application.

Method	AR				FERET			
	cost	TER (%)	FRR (%)	FAR (%)	cost	TER (%)	FRR (%)	FAR (%)
PCA	557	13.21	23.33	7.14	329	10.18	20.95	3.71
CSPCA	482	12.32	22.86	6.00	272	9.64	20.95	2.86
LDA	275	6.07	10.48	3.43	129	6.79	16.67	0.86
CSLDA	199	5.71	11.43	2.29	100	5.89	14.76	0.57
LPP	853	21.25	38.57	10.86	331	10.36	21.43	3.71
CSLPP	591	15.36	20.57	8.43	202	6.07	12.38	2.29

the training set and testing set each includes N_G images with M gallery subjects and N_I images as impostors randomly selected from the entire database¹. We repeated this selection 10 times for each database and computed the average recognition accuracy as the final recognition result. We have also compared the proposed method with the three conventional subspace learning methods, such as PCA [1], LDA [2], and LPP [8]. We applied the nearest neighbor (NN) rule in the Euclidean space to perform recognition. We compared the total cost (cost), total error rate (TER), false rejection rate (FRR), and false acceptance rate (FAR) of the methods. To provide a fair comparison, all the results reported here are based on the best tuned parameters of all the methods compared. The results are tabulated in Table. 5.6. We can see from this table that our proposed cost-sensitive subspace learning methods have smaller total cost than the existing, uniform-cost subspace learning methods.

We also investigated the performance of the proposed cost-sensitive methods versus different numbers of gallery subjects (M) on the AR and FERET databases. For the AR database, M varies from 10 to 60 with increments of 10, and for the FERET database, M varies from 10 to 70 with increments of 10. Fig. 5.4 shows the total costs of each method

¹Note that the imposters are from different sets of subjects in our experiments.

Fig. 5.4: Comparison on total cost with different number of gallery subjects. The first row are the results on the AR database. The second row are the results on the FERET database.

obtained on the AR and FERET databases, respectively. We can clearly see from this figure that our proposed cost-sensitive methods perform better than the existing uniform-cost methods under different numbers of gallery subjects in terms of the total cost.

We further conducted experiments on the AR database to compare the performance of the proposed cost-sensitive methods using different classifiers. The values of parameters M , N_G , N_I , C_{IG} , C_{GI} and C_{GG} were set to those as listed in Table 5.5. We selected the latest cost-sensitive classifier, i.e., the $mckNN$ classifier [36] and compared the performance with that of the conventional kNN classifier. We set $k = 1$ in our experiments and hence the conventional kNN classifier is the well-known NN rule. Fig. 5.5 shows the total costs of CSPCA, CSLDA and CSLPP, respectively, when the two different classifiers were applied on the AR database. We can see from this figure that the total costs can be further reduced when the cost-sensitive classifier was applied.

Lastly, we examined the computational costs of the proposed cost-sensitive methods and compared them with the corresponding uniform-cost methods including PCA, LDA and LPP. The computer used for conducting the experiments comprises a 2.66-GHz CPU and 2G RAM. Table 5.7 shows the CPU time spent on the training and the testing (recognition) phases by these subspace learning methods using Matlab software and the

Fig. 5.5: Total cost on the AR database of different cost-sensitive subspace learning methods when the uniform-cost and cost-sensitive classifiers are applied, respectively.

Table 5.7: CPU time (in second) used by different methods on the AR database.

Method	Training	Test
PCA	0.1342	0.0029
CSPCA	0.1463	0.0029
LDA	0.1563	0.0022
CSLDA	0.2188	0.0022
LPP	0.3593	0.0024
CSLPP	0.4012	0.0024

AR database. Note that the values of parameters M , N_G , N_I , C_{IG} , C_{GI} and C_{GG} were shown in Table. 5.5 and the classifier used was the NN classifier. We can see that the training and testing computational cost of our proposed cost-sensitive methods are comparable to those of the uniform-cost methods.

Application 2—Criminal Search: Besides access control, face recognition can be also used to other security applications. A representative example is face recognition-based criminal search. The aim of the such system is to search and identify the criminal suspects from a large face database. Therefore, the gallery subjects are the criminal suspects and the impostors are the innocent citizens, respectively, for this application. Hence, $C_{GI} > C_{GG} > C_{IG}$. We set the parameters as shown in Table. 5.8 to simulate the real conditions,

Table 5.8: Parameter settings for a face recognition-based criminal search system.

Database	M	N_G	N_I	$C_{GI} : C_{GG} : C_{IG}$
AR	30	7	350	100:10:1
FERET	70	3	350	100:10:1

Table 5.9: Comparison of total cost (cost), total error rate (TER), false rejection rate (FRR) and false acceptance rate (FAR) on the AR and FERET databases for the criminal search application.

Method	AR				FERET			
	cost	TER (%)	FRR (%)	FAR (%)	cost	TER (%)	FRR (%)	FAR (%)
PCA	2316	13.21	23.33	7.14	1569	10.18	20.95	3.71
CSPCA	2154	11.79	21.90	5.71	1296	9.64	20.95	2.86
LDA	1339	6.07	10.48	3.43	624	6.79	16.67	0.86
CSLDA	1171	6.61	12.86	2.86	364	5.54	14.29	0.29
LPP	3969	21.25	38.57	10.86	975	10.36	21.43	3.71
CSLPP	2759	15.36	28.57	7.43	970	6.07	12.38	2.29

and conducted face recognition on the AR and FERET databases, respectively. Similar to the pervious access control application, we have also compared the proposed method with the three conventional subspace learning methods, such as PCA [1], LDA [2], and LPP [3], and applied the nearest neighbor (NN) rule in the Euclidean space to perform recognition. The results are tabulated in Table. 5.9. Similar to previous face recognition experiments, we can also see from this table that our proposed cost-sensitive subspace learning methods have smaller total cost than the existing, uniform-cost subspace learning methods.

We also investigated the performance of the proposed cost-sensitive methods versus different numbers of gallery subjects (M) on the AR and FERET databases. For the AR database, M varies from 10 to 60 with increments of 10, and for the FERET database, M varies from 10 to 70 with increments of 10. Fig. 5.6 shows the total costs of each method obtained on the AR and FERET databases, respectively. We can clearly see from this figure that our proposed cost-sensitive methods performs better than the existing uniform-cost methods under different numbers of gallery subjects in terms of the total cost.

Fig. 5.6: Comparison on total cost with different number of gallery subjects. The first row are the results on the AR database. The second row are the results on the FERET database.

5.3.3 Discussion

The following two observations can be made from the experimental results listed in Tables 5.3-59 and Figs. 5.3-5.6:

1. Our proposed cost-sensitive methods can achieve higher recognition performance than the existing uniform-cost methods in terms of the total cost in all face recognition experiments. For example, the least cost gain of our cost-sensitive methods over the corresponding uniform-cost methods on the AR and FERET databases are 76 and 71 for the access control, and 162 and 5 for the criminal search applications, respectively. The reason is our cost-sensitive approach aims to seek a minimal loss (cost) and existing uniform-cost methods aims to seek a minimal recognition error.

2. Our proposed cost-sensitive methods sometimes performs better the existing uniform-cost methods in terms of the total error rate in some face recognition experiments. For example, our proposed cost-sensitive methods consistently outperform the corresponding uniform-cost methods in terms of the total error rate on the access control example. It looks a little strange at first glance because it seems to violate the objective of cost-sensitive approach. The results are reasonable because compared with PCA and

LPP which are unsupervised, our proposed CSPCA and CSLPP are supervised, and supervised methods are generally outperform unsupervised methods for recognition tasks. Hence, CSPCA and CSLPP can obtain less recognition error rates than PCA and LPP in some cases. Even if LDA and CSLDA are both supervised, CSLDA also can outperform LDA in some recognition experiments and that is because even LDA aim to seek low recognition error, it has not considered the different importance of the samples in the training set, especially those from the impostor class as these samples are much easier to be misclassified. However, CSLDA emphasizes different samples in the learning phase and possibly obtains better recognition accuracy than LDA.

5.4 Chapter Summary

We have proposed in this chapter a cost-sensitive subspace learning approach for face recognition. Motivated by the fact that different types of misclassifications in a real-world face recognition system usually lead to different losses, we have presented three cost-sensitive subspace learning methods, namely, cost-sensitive principal component analysis (CSPCA), cost-sensitive linear discriminant analysis (CSLDA), and cost-sensitive locality preserving projections (CSLPP), such that low recognition losses rather than low recognition errors can be attained when performing recognition in the low-dimensional subspaces derived. Experimental results have demonstrated the efficacy of the proposed approach. How to learn an appropriate cost matrix for a specific real-world face recognition application and extend our cost-sensitive approach to other subspace learning methods appear to be interesting directions of future work. Partial work presented in this chapter was published in [164].

Chapter 6

Subspace Learning with Limited Number of Training Samples

6.1 Introduction

While many existing subspace methods have achieved reasonably good performance in pattern classification, their performance degrade much when the number of training samples per class is small. This is because a limited number of training samples are usually not able to effectively characterize the geometrical structure of the original sample space. In many practical pattern classification applications, such as face recognition, it is difficult and expensive to collect a large number of training samples, and we need to fully exploit the discriminant information from the available limited training samples. Aiming at generalizing the representative capacity of the training samples, Li and Lu [70] proposed a novel nearest feature line (NFL) method to virtually enlarge the training set for face recognition. The basic idea of NFL is using a linear model to interpolate and extrapolate each pair of prototypes belonging to the same class to model the possible variants of the training samples. Because of this excellent generalization capacity, NFL has been widely used in many computer vision and pattern recognition applications, such as face recognition [70, 74, 75], audio classification [71, 72, 76], and image retrieval [72].

The original NFL method is, however, computationally expensive because each pair of the intraclass samples is utilized to construct a feature line (FL) to enlarge the training set. To address this shortcoming, Zheng *et al.* [77] proposed a nearest neighbor line (NNL) method by selecting the nearest FL of each class for classification; Gao *et al.* [78]

proposed a nearest center-based line (NCL) method by constructing a center-based line (CL) instead of FL to virtually enlarge the training set; Du *et al.* [79] proposed a rectified nearest feature line segment (RNFLS) method by deleting unimportant FLs to improve the efficiency of NFL. While these methods can reduce the high computational cost of the original NFL method, all of them apply FLs to virtually enlarge the training set and may not fully depict the geometrical information of the training samples since the variants of the samples in each class may not effectively be characterized by these FLs.

On the other hand, Chien and Wu [84] proposed a novel nearest feature space (NFS) method, where the feature space (FS) of each class is constructed in the manner similar to the FL constructed in NFL to model the variants of the training samples. There are two notable merits of NFS compared with NFL: NFS needs constructing only one FS for each class of the training samples, which can significantly reduce the computational complexity of the original NFL method; furthermore, NFS can better handle the variations of the samples than NFL as it can provide more information to virtually the training samples. Empirical results in [19] have clearly demonstrated these advantages.

However, most of the existing NFL- and NFS-based methods [70–79] use only the NFL and NFS metrics in the classification phase and not in the feature extraction (subspace learning) phase. While the classification performance can be enhanced by NFL and NFS to a certain extent, the learning ability of existing subspace learning methods remains to be poor when the number of training samples is limited. To enhance the discriminative ability of the feature subspace, some recent work has utilized the NFL metric in subspace learning for feature extraction. For example, Zheng *et al.* [80] presented a nearest feature line-based nonparametric discriminant analysis (NFL-NDA) algorithm, and Pang *et al.* [81] put forward a nearest feature line-based space (NFLS) method, respectively. These recently proposed NFL-based methods have made use of the NFL metric in both the learning and classification phases, and have attained improved recognition performance. As the subspace was learned with the use of FL distances, these methods have mitigated the limitation arising from the small number of training samples, and perform better than conventional subspace learning methods such as PCA and LDA, which learn a subspace by using the conventional NN metric.

In this chapter, we propose applying NFL and NFS metrics to learn a feature subspace for feature extraction. Since existing subspace learning algorithms degrade heavily when

the number of training samples per class is small because such limited samples are usually not able to effectively characterize the geometrical structure of the original sample space for learning an effective feature subspace, we propose using the nearest feature line (NFL) and nearest feature space (NFS) metrics rather than the conventional nearest neighbor (NN) metric to seek discriminative feature spaces for feature extraction. Experimental results are presented to demonstrate the efficacy of the proposed methods.

The rest of this chapter is organized as follows. Section 6.2 reviews the exiting NFL and NFS methods. Section 6.3 details the proposed methods. Section 6.4 presents the experimental results, and Section 6.5 concludes this chapter.

6.2 NFL and NFS

6.2.1 NFL

Given two intraclass training samples x_1 and x_2 , a straight line passing through them, denoted as $\overline{x_1x_2}$ as shown in Fig. 6.1, is called a FL of the class that they belong to. It is easy to see that $\overline{x_1x_2}$ provides not only possible variants of x_1 and x_2 , but also more virtual samples of the same class for training. The NFL distance between a probe sample x_p and the FL $\overline{x_1x_2}$ is defined as the Euclidean distance between the probe sample x_p and its projection on FL $\overline{x_1x_2}$, x_q , as shown in Fig. 6.1 and given below [70]:

$$d(x_p, \overline{x_1x_2}) = \|x_p - x_q\| \quad (6.1)$$

where $\overline{x_px_q}$ is perpendicular to $\overline{x_1x_2}$, $x_q = x_1 + \mu(x_2 - x_1)$, and $\mu = \frac{(x_p - x_1)^T(x_2 - x_1)}{(x_2 - x_1)^T(x_2 - x_1)}$. Obviously, when $\mu = 0$, $x_q = x_1$, when $\mu = 1$, $x_q = x_2$, and when $0 < \mu < 1$, x_q is a linear interpolation between x_1 and x_2 . Furthermore, when $\mu < 0$ or $\mu > 1$, x_q is a linear extrapolation of x_1 and x_2 . See [70] for more details.

6.2.2 NFS

Let $x = [x_{k1}, x_{k2}, \dots, x_{kt}]$ be the independent prototypes in the k th class of the training set. The space $S^k = sp(x_{k1}, x_{k2}, \dots, x_{kt})$ is called the FS of the k th class, where $sp(a_1, a_2, \dots, a_k)$ represents the span of the set $\{a_1, a_2, \dots, a_k\}$ [84]. It is obvious that this feature space contains more variants of the samples in this class than the FL in the

Fig. 6.1: The feature line $\overline{x_1x_2}$ of two samples, a probe sample x_p , and its projection on the feature line, x_q .

NFL method. For a probe sample x_p , the FS distance between x_p and the space S^k , is defined as the Euclidean distance between the probe sample x_p and its projection on S^k , x_q , given below [84]:

$$\begin{aligned} d(x_p, S^k) &= \|x_p - x_q\| \\ &= \left\| x_p - \sum_{i=1}^t w_{ki} x_{ki} \right\|^2 \end{aligned} \quad (6.2)$$

where w_{ki} is the coefficient of x_{ki} , $\sum_{i=1}^t w_{ki} = 1$, and its closed-form solution can be obtained according to the scheme developed in locally linear embedding (LLE) [6].

Fig. 6.2 shows an example of the FS distance between a probe sample x_p and a feature space S^k , spanned by three independent samples x_{k1} , x_{k2} and x_{k3} . We can see from this figure that NFS can virtually have more samples than the NFL methods.

6.3 Proposed Methods

6.3.1 DNFLA

As mentioned in Section I, NFLS minimizes only the FL distances of within-class samples and does not consider the FL distances of between-class samples. However, minimizing the within-class NFL distances may not lead to an effective discriminant subspace for recognition task. Based on this motivation, we formulate the proposed DNFLA method

Fig. 6.2: The feature space distance between a probe sample x_p , and its projection on the feature space, x_q .

as the following optimization problem:

$$\min_W J(W) = \frac{1}{NI_i} \sum_{i=1}^N \sum_{j \in P_i} \|W^T x_i - W^T x_{ij}\|^2 - \frac{1}{NE_i} \sum_{i=1}^N \sum_{k \in R_i} \|W^T x_i - W^T x_{ik}\|^2 \quad (6.3)$$

where $P_i = \{j | l_{x_j} = l_{x_i}\}$, $R_i = \{k | l_{x_k} \neq l_{x_i}\}$, x_{ij} and x_{ik} denote the projection of x_i onto the FLs formed by the samples shared the same and different labels with x_i , respectively, I_i and E_i are the numbers of samples in P_i and R_i .

With some algebraic deduction, we can simplify $J(W)$ to the following form

$$\begin{aligned} J(W) &= \frac{1}{NI_i} \sum_{i=1}^N \sum_{j \in P_i} \|W^T x_i - W^T x_{ij}\|^2 - \\ &\quad \frac{1}{NE_i} \sum_{i=1}^N \sum_{k \in R_i} \|W^T x_i - W^T x_{ik}\|^2 \\ &= \frac{1}{NI_i} \sum_{i=1}^N \sum_{j \in P_i} \text{tr}[W^T (x_i - x_{ij})(x_i - x_{ij})^T W] \\ &\quad - \frac{1}{NE_i} \sum_{i=1}^N \sum_{k \in R_i} \text{tr}[W^T (x_i - x_{ik})(x_i - x_{ik})^T W] \\ &= \text{tr}\{W^T [\frac{1}{NI_i} \sum_{i=1}^N \sum_{j \in P_i} (x_i - x_{ij})(x_i - x_{ij})^T] W\} \\ &\quad - \text{tr}\{W^T [\frac{1}{NE_i} \sum_{i=1}^N \sum_{k \in R_i} (x_i - x_{ik})(x_i - x_{ik})^T] W\} \\ &= \text{tr}[W^T (A - B)W] \end{aligned} \quad (6.4)$$

where $A \triangleq \frac{1}{NI_i} \sum_{i=1}^N \sum_{j \in P_i} (x_i - x_{ij})(x_i - x_{ij})^T$ and $B \triangleq \frac{1}{NE_i} \sum_{i=1}^N \sum_{k \in R_i} (x_i - x_{ik})(x_i - x_{ik})^T$.

We then impose an constraint on $w^T w = 1$ to make the objective function to be well-posed respective to w , and the bases of DNFLA can be obtained by solving the following eigenvalue problem

$$(A - B)w = \lambda w \quad (6.5)$$

Let $w_1, w_2, \dots, w_q$ be the eigenvectors of (6.5) corresponding to the q smallest eigenvalues ordered according to $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_q$ ¹. A $d \times q$ transformation matrix $W = [w_1, w_2, \dots, w_q]$ can be obtained to project each sample $d \times 1$ x_i into a feature vector $q \times 1$ y_i , as follows:

$$y_i = W^T x_i, \quad i = 1, 2, \dots, N \quad (6.6)$$

6.3.2 NFSA

While the original NFS method has demonstrated better classification performance than the conventional NN and NFL methods [84], most conventional subspace learning methods only utilize the NFS metric in the classification phase and not in the feature extraction (subspace learning) phase. For pattern classification tasks, it is desirable to utilize the NFS metric in both the feature extraction and classification phases to improve the final recognition performance. Motivated by this observation, we propose a new subspace learning method, called nearest feature space analysis (NFSA), by using the NFS metric for feature extraction.

Assume that the training set is represented in a matrix form $X = [x_1, x_2, \dots, x_N]$, where $x \in R^d$ and N is the number of training samples. The corresponding class label of x_i is denoted as $l_{x_i} \in \{1, 2, \dots, c\}$, where $i = 1, 2, \dots, N$ and c is the total number of classes. Let N_k be the number of samples in the k th class, $k = \{1, 2, \dots, c\}$, then $N = \sum_{k=1}^c N_k$. We assume N_k is usually larger than 3 as it is the case in most classification

¹Since there could be some problem if the difference of the two matrices is negative and some eigenvalues could be negative. For this case, the negative eigenvalues should be removed and the smallest nonnegative eigenvalues are selected and the corresponding eigenvectors are chosen to construct the low-dimensional feature subspace.

tasks. Hence, the aim of NFSA is to seek a projection W that maps x_i to a low-dimensional subspace by $y_i = W^T x_i$ such that the sum of within-class FS distances, defined as

$$\min_W J(W) = \sum_{i=1}^N \|W^T x_i - W^T x_i^{fs}\|^2 \quad (6.7)$$

is minimized, where x_i^{fs} denotes the projection of x_i onto the FS formed by the other samples shared the same class label with x_i .

By simple algebraic deduction, the objective of NFSA can be rewritten as

$$\begin{aligned} J(W) &= \sum_{i=1}^N \|W^T x_i - W^T x_i^{fs}\|^2 \\ &= \text{tr}\{W^T [\sum_{i=1}^N (x_i - x_i^{fs})(x_i - x_i^{fs})^T] W\} \\ &= \text{tr}(W^T H W) \end{aligned} \quad (6.8)$$

where $H \triangleq \sum_{i=1}^N [(x_i - x_i^{fs})(x_i - x_i^{fs})^T]$.

We then impose an constraint on $w^T w = 1$ to make the objective function to be well-posed respective to w , and the bases of NFSA can be obtained by solving the following eigenvalue problem

$$Hw = \lambda w \quad (6.9)$$

Let $w_1, w_2, \dots, w_m$ be the eigenvectors resulted from (6.9), ordered according to their eigenvalues $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$ ², then $W = [w_1, w_2, \dots, w_m]$ is the mapping of NFSA.

6.3.3 DNFSa

NFSA minimizes only the FS distances of within-class samples and does not consider the FS distances of between-class samples. However, minimizing the within-class FS distances may not lead to an effective discriminant subspace for recognition tasks. Fig. 6.3 provides a simple example with two classes of samples denoted by squares and circles to illustrate

²Similar to DNFLA, there could also be some problem if the difference of the two matrices is negative and the negative eigenvalues should be removed and the smallest nonnegative eigenvalues are selected and the corresponding eigenvectors are chosen to construct the low-dimensional feature subspace.

Fig. 6.3: A simple example illustrating the subspace projection by the NFS and DNFS methods: (a) Original data, (b) NFS, (c) DNFS.

this point. Fig. 6.3(a) shows the geometrical structure of the original data. The square samples (x_{s1} , x_{s2} , x_{s3} and x_p) belong to the first class and the circle samples (x_{t1} , x_{t2} and x_{t3}) belong to the second class, where x_{s1} , x_{s2} , x_{s3} , x_{t1} , x_{t2} and x_{t3} are the training samples and x_{t3} is the testing sample. The projection point of the testing sample x_p on the feature space of the square and circle classes are x_t and x_s , respectively. Hence, the Euclidean distance between x_p and x_t is the FS distance between the testing sample and the square class, and that between x_p and x_s is the FS distance between the testing sample and the circle class. Suppose that the Euclidean distances between x_p and x_t is equal to that between x_p and x_s , and the testing sample x_p cannot be correctly classified by the NFS metric. Without loss of generality, the samples in Fig. 6.3(a) projected by NFS are shown in Fig. 6.3(b), where the within-class FS distances decrease but the Euclidean distances between x_p and the two feature spaces remain to be equal, so do the FS distances and the samples in the two classes. Hence, NFS contributes little for the classification of x_p . A desirable NFS-based subspace learning method should maximize the between-class FS distances and minimize the within-class FS distances, simultaneously. As shown in Fig. 6.3(c), the FS distances between x_p and the circle class are expected to be minimized and those between x_p and the square class are to be maximized in the reduced subspace. Thus x_p can be correctly classified into the circle class.

Based on the above motivation, we formulate the proposed DNFSFA method as the following optimization problem:

$$\begin{aligned} \max_W J(W) = & \frac{1}{N(c-1)} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq l(x_i)}}^c \|W^T x_i - W^T x_{ij}^{inter}\|^2 \\ & - \frac{1}{N} \sum_{i=1}^N \|W^T x_i - W^T x_i^{intra}\|^2 \end{aligned} \quad (6.10)$$

where x_{ij}^{intra} denotes the projection of x_i onto the FS formed by the samples shared the same label with x_i , and x_{ij}^{inter} denotes the projection of x_i onto the FS formed by the j th class of samples ($j \neq i$), respectively.

With some algebraic deduction, we can simplify $J(W)$ to the following form

$$\begin{aligned} J(W) = & \frac{1}{N(c-1)} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq l(x_i)}}^c \|W^T x_i - W^T x_{ij}^{inter}\|^2 \\ & - \frac{1}{N} \sum_{i=1}^N \|W^T x_i - W^T x_i^{intra}\|^2 \\ = & \frac{1}{N(c-1)} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq l(x_i)}}^c \text{tr}[W^T (x_i - x_{ij}^{inter})(x_i - x_{ij}^{inter})^T W] \\ & - \frac{1}{N} \sum_{i=1}^N \text{tr}[W^T (x_i - x_i^{intra})(x_i - x_i^{intra})^T W] \\ = & \text{tr}\{W^T [\frac{1}{N(c-1)} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq l(x_i)}}^c (x_i - x_{ij}^{inter})(x_i - x_{ij}^{inter})^T] W\} \\ & - \text{tr}\{W^T [\frac{1}{N} \sum_{i=1}^N (x_i - x_i^{intra})(x_i - x_i^{intra})^T] W\} \\ = & \text{tr}[W^T (S_1 - S_2) W] \end{aligned} \quad (6.11)$$

where $S_1 \triangleq \frac{1}{N(c-1)} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq l(x_i)}}^c (x_i - x_{ij}^{inter})(x_i - x_{ij}^{inter})^T$, and $S_2 \triangleq \frac{1}{N} \sum_{i=1}^N (x_i - x_i^{intra})(x_i - x_i^{intra})^T$.

We also impose a length constraint $w^T w = 1$ on the proposed DNFSFA. Then, the projection W of DNFSFA can be obtained by solving the following eigenvalue problem

$$(S_1 - S_2)w = \lambda w. \quad (6.12)$$

Let $w_1, w_2, \dots, w_q$ be the eigenvectors of (6.12) corresponding to the q largest eigenvalues ordered according to $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$. A $d \times q$ transformation matrix $W = [w_1, w_2, \dots, w_q]$ can be obtained to project each sample $d \times 1$ x_i into a feature vector $q \times 1$ y_i , as follows:

$$y_i = W^T x_i, \quad i = 1, 2, \dots, N \quad (6.13)$$

6.4 Experimental Results

In this section, we evaluate the effectiveness of our proposed methods for image classification, including face recognition and facial expression recognition. We also compare the proposed methods with some conventional NN-based subspace learning methods, such as PCA [1], LDA [2], LPP [8], and MFA [17], and the latest NFLS [81]. To have a fair comparison, all the results reported here are based on the best tuned parameters of all the compared methods.

6.4.1 Face Recognition

We conducted face recognition experiments on the widely used AR [109] face databases. The AR database contains over 4000 face images of 126 subjects (70 men and 56 women). We selected a subset of 826 face images of 118 subjects with 7 face images of different expressions and lighting conditions per subject. The facial part of each image was manually cropped into 64×64 pixels and aligned according to the eyes' positions. We randomly selected 4 images of each subject to construct the training set and used the remaining to form the testing set. To incorporate the variations in the classification results, we computed the average recognition accuracy (ARA) of 20 runs with different training and testing sets. Table 6.1 tabulates the ARA attained by each method with the corresponding feature dimensions indicated in the parentheses. The ARAs versus feature dimensions of different subspace learning methods are shown in Fig. 6.4, where the results obtained by employing the nearest neighbor (NN) classifier, the nearest feature line (NFL) classifier, and the nearest feature space (NFS) classifier are shown in Fig. 6.4(a), Fig. 6.4(b), and Fig. 6.4(c), respectively.

Table 6.1: Comparison on ARA (%) of different algorithms using the AR database.

Method	NN	NFL	NFS
PCA	80.37 (60)	82.37 (60)	84.37 (60)
LDA	89.26 (40)	91.26 (40)	93.86 (60)
LPP	91.29 (80)	92.28 (80)	94.56 (60)
MFA	91.65 (80)	92.55 (80)	95.15 (60)
NFLS	92.21 (60)	93.20 (60)	95.21 (60)
DNFLA	93.64 (60)	94.62 (60)	95.64 (60)
NFSA	94.88 (60)	95.86 (60)	95.88 (60)
DNFSA	95.68 (60)	96.88 (60)	96.32 (60)

As can be seen from Table 6.1, the proposed DNFSA (NFSA) methods outperform PCA, LDA, LPP, MFA, NFLS, and DNFLA with gains in accuracy of 15.31% (14.51%), 6.42% (5.62%), 4.39% (3.59%), 4.03% (3.23%), 3.47% (2.67%), and 2.04% (1.24%) when applying the nearest neighbor (NN) classifier, and 14.51% (13.49%), 5.62% (4.60%), 4.60% (3.58%), 4.33% (3.29%), 3.68% (2.66%), and 2.26% (1.24%) when using the nearest feature line (NFL) classifier, and 11.95% (11.51%), 2.46% (2.02%), 1.76% (1.32%), 1.17% (0.73%), 1.11% (0.67%), and 0.68% (0.22%) when employing the nearest feature space (NFS) classifier, respectively. From Fig. 6.4, we can see our proposed NFSA and DNFSA methods achieve higher recognition accuracy than the other compared methods in terms of the ARAs.

6.4.2 Facial Expression Recognition

We performed facial expression recognition on a widely used facial expression databases, namely, the Cohn-Kanade (C-K) [110] face database. We selected 10 subjects who displayed all six basic facial expressions, namely, anger, disgust, fear, happiness, sadness, and surprise to construct the dataset. Some sample images of one subject with different facial expressions are shown in Fig. 6.5. Each subject from the C-K database forms an expression over time starting from a neutral pose and ending with a very intense expression, thus having several video frames with different expression intensities. However, the number of these intermediate video frames is not the same for different subjects. We have selected the last two frames of each video sequence with high facial expression intensity, and used them to form the database utilized in our experiments. The total number of

(a)

(b)

(c)

Fig. 6.4: ARAs achieved by various methods under different feature dimensions. The results obtained when (a) the NN classifier, (b) the NFL classifier, and (c) the NFS classifier are applied, respectively.

frontal images used was 120. The facial part of each image was manually cropped into 64×64 pixels and aligned according to the eyes' positions.

For each facial expression, we randomly selected 10 images to construct the training set and used the remaining to form the testing set. To incorporate the variations in the classification results, we computed the average recognition accuracy (ARA) of 20 runs with different training and testing sets similar to face recognition discussed before. Table 6.2 tabulates the ARA attained by each method with the corresponding feature

Fig. 6.5: Samples of face images of a subject from the C-K database with different expressions. (a) Surprise. (b) Fear. (c) Sadness. (d) Anger, (e) Disgust, and (f) Happy.

dimensions indicated in the parentheses. The ARAs versus feature dimensions of different subspace learning methods are shown in Fig. 6.6, where the results obtained by employing the nearest neighbor (NN) classifier, the nearest feature line (NFL) classifier, and the nearest feature space (NFS) classifier are shown in Fig. 6.6(a), Fig. 6.6(b), and Fig. 6.6(c), respectively.

As can be seen from Table 6.2, the proposed DNFS (NFS) methods outperforms PCA, LDA, LPP, MFA, NFLS, and DNFLA with gains in accuracy of 15.66% (14.55%), 5.25% (4.14%), 4.66% (3.55%), 2.83% (1.72%), 2.66% (1.55%), and 2.00% (0.89%) when applying the nearest neighbor (NN) classifier, 14.28% (12.72%), 4.08% (2.52%), 3.95% (2.39%), 2.62% (1.06%), 2.08% (0.52%), and 1.63% (0.07%) when using the NFL classifier, and 9.91% (9.34%), 3.57% (3.00%), 3.12% (2.55%), 1.91% (1.34%), 1.57% (1.00%), and 1.12% (0.55%) when employing the NFS classifier, respectively. We can see from Fig. 6.6 our proposed NFS and DNFS methods consistently achieve higher recognition

Table 6.2: Comparison on ARA (%) of different algorithms using the C-K database.

Method	NN	NFL	NFS
PCA	76.67 (6)	78.67 (7)	83.33 (6)
LDA	87.08 (5)	88.87 (5)	89.67 (5)
LPP	88.67 (8)	89.00 (8)	90.12 (7)
MFA	89.50 (5)	90.33 (5)	91.33 (5)
NFLS	89.67 (7)	90.87 (8)	91.67 (7)
DNFSA	90.33 (8)	91.32 (7)	92.12 (8)
NFSA	91.22 (7)	91.39 (7)	92.67 (7)
DNFSA	92.33 (7)	92.95 (7)	93.24 (7)

Table 6.3: CPU time (in second) used by different methods on the AR database.

Method	PCA	LDA	LPP	MFA	NFLS	DNFLA	NFLS	DNFSA
Training	0.8125	1.1719	1.1406	1.9375	8.7656	7169.6250	1.4375	171.0625
Recognition	0.0018	0.0018	0.0018	0.0018	0.0018	0.0018	0.0018	0.0018

accuracy than the compared methods in terms of the ARAs.

6.4.3 Computational Complexity

Lastly, we examined the computational costs of the proposed methods and compared them with those of other subspace learning methods including PCA, LDA, LPP, MFA, NFLS and DNFLA. Our hardware configuration comprises a 2.66-GHz CPU and 2GB RAM. Table 6.3 shows the CPU times spent on the training and the testing (recognition) phases by these subspace learning methods using Matlab software and the AR database with the nearest neighbor (NN) classifier. Note that 4 samples of each subject on this database were selected for training and the dimension of all the reduced subspaces was set to the same (117).

From Table 6.3, we can see that the computational complexity of the proposed NFL- and NFS-based subspace learning algorithms for training are generally larger than the conventional NN-based subspace learning methods, such as PCA, LDA, LPP and MFA, and smaller than the newly proposed NFL-based subspace learning methods, such as NFLS and DNFSA. In practical applications, however, training is usually an offline process and only recognition needs to be performed in real time. Thus, the recognition time is usually more of our concern than the training time. As shown in Tables 6.1-6.2, the

recognition time of the proposed algorithms is the same as other approaches for the same number of features. Moreover, the proposed algorithms can potentially be more computationally efficient for recognition than other approaches including PCA, LDA, LPP, MFA, NFLS and DNFLA because our methods require a smaller number of features than other approaches to obtain the same recognition accuracy, as shown in Fig. 6.4 and Fig. 6.6.

6.4.4 Discussion

We make three main observations from the experimental results listed in Tables 6.1-6.2 and Fig. 6.4 and Fig. 6.6:

1. NFL- and NFS-based subspace learning methods achieve higher recognition accuracy than the conventional NN-based subspace learning methods including PCA, LDA, LPP and MFA in terms of recognition accuracy. This is because both NFL and NFS depict the possible variants of samples of each training class and virtually enlarge the training set, and the FL and FS metrics can better characterize the distance between one sample and other samples, especially when the number of training samples per class is small.
2. NFS-based subspace learning methods achieve higher recognition accuracy than the newly proposed NFL-based subspace learning methods including NFLS and DNFLA. The improvement is due to that the FS metric can generate more training samples to better characterize the geometrical information of the samples than the FL metric.
3. All subspace learning methods under comparison attain better classification performance when the NFL and NFS classifiers are applied. This further implies that the FL and FS distances can better characterize the geometrical structure of the samples than the conventional Euclidean distance used in the NN classifier.

6.5 Chapter Summary

We have proposed in this chapter three new subspace learning methods, called discriminant nearest feature line analysis (DNFLA), nearest feature space analysis (NFSA) and discriminant nearest feature space analysis (DNFSA) for feature extraction and classification by using the nearest feature line (NFL) and nearest feature space (NFS) metrics,

respectively. Compared with the conventional NN-based subspace learning algorithms such as PCA, LDA, LPP and MFA, the proposed methods can better characterize the geometrical information of the training samples by inheriting the merit of NFL and NFS. Experimental results on face and facial expression recognition have demonstrated the efficacy of the proposed methods. Partial work presented in this chapter were published in [166] and [167].

(a)

(b)

(c)

Fig. 6.6: ARAs achieved by various methods under different feature dimensions. The results obtained when (a) the NN classifier, (b) the NFL classifier, and (c) the NFS classifier are applied, respectively.

Chapter 7

Multi-Label Subspace Learning for Human Age Estimation

7.1 Introduction

Human gait is a useful biometric signature and it has received much attention in the computer vision community. The interest is strongly driven by the need for intelligent surveillance in security-sensitive environments such as banks, airports, and parking lots. Existing human gait analysis mainly focuses on gait-based human identification at a distance, and many methods have been proposed in the literature [111, 131–141]. While these methods have attained reasonably good performance when the training and test data were acquired under similar conditions, they require the gait sequences of the subjects concerned to be acquired in advance for training purpose. In other words, if a person’s gait samples are not pre-stored in the systems, the identity would not be recognized by these gait recognition methods.

However, in some open visual surveillance systems such as airports, railway stations and building entrances, there are potentially many unregistered persons, and it’s extremely difficult or even impossible to collect all their gait samples in advance for identification purpose. Under such circumstance, recognizing other biometric attribution, such as age information, of the persons concerned may also be useful and desirable.

In addition to identity, age is also an important attribute of a person. However, existing age estimation methods mainly rely on human facial image analysis [85–103]. In controlled visual surveillance, there is usually some distance between the camera and

the person of interest, and the facial images collected under such scenarios are usually of low-resolution or with limited details. Analyzing such facial images may not lead to accurate age estimation. However, human gait can be well perceived under these scenarios, especially at a distance. Consequently it can be used for age estimation and has high application potentials. For example, it can be used in entrance monitoring to set off an alert if the estimated age is forbidden. It can also be used to count the number of people under different age groups entering into a region of interest to enhance the intelligence of a visual surveillance system. Hence, gait-based age estimation is of great interest and significance. However, to our knowledge, there is not any prior work focusing on using gait signatures to estimate human ages at a distance.

An age estimation system usually consists of two modules [85, 86]: feature representation and age estimation. Existing facial feature representation methods for human age estimation mainly include anthropometric model [87], active appearance model (AAM) [101, 102], aging pattern subspace [92–94], aging manifold analysis [89–91], part-based [96] and patch-based [99] appearance model, local binary patterns (LBP) [95], and bio-inspired features (BIF) [85, 86]. Given a feature representation, age estimation can be implemented by using a multi-class classification [87] or regression [89, 90] method. While these methods have achieved reasonable performance, most of them are arguably not suitable to be extended for gait-based human age estimation for the following reasons:

- 1) The cranio-facial development theory and facial skin wrinkle analysis are used to create the anthropometric model [87], in which the changes of shape and texture patterns are measured to categorize a face into several age groups. AAM is suitable for facial image modeling [100, 102] as it can well model the shape and texture variations to reflect the aging of facial images. These methods, however, are not suitable for gait-based age estimation as there is no apparent shape and texture changes of human gait patterns when a person ages.

- 2) To handle incomplete data such as missing ages in the training data and to model the manifold of aging patterns, the aging pattern subspace models a sequence of individual aging patterns by using a subspace representation [92–94], and the aging manifold analysis method depicts the manifold structure of human aging to estimate the human age [89–91]. However, both of them usually need many facial images of a certain person under different

Fig. 7.1: Gait sequences of two persons with the same age (26 years old) but different gender. (a) Male and (b) Female.

ages to model the aging manifold. In real-world visual surveillance applications, it's generally not easy to collect a large number of gait sequences of a certain person under different ages. Hence, the aging subspace and manifold analysis methods are also not suitable for gait-based age estimation.

3) For the regression methods [89–91, 100], some face images are firstly applied to learn the regression model, which is then used to estimate the age from a given facial image. Even they have achieved reasonably good performance and are popular for age estimation, it is still an open problem to design an effective regression function to precisely model the human aging process, in particular from facial images of low resolution.

In addition to the above-mentioned work on facial image-based age estimation, there have been some attempts on gait-based human gender and age analysis in the literature. For example, the discussions in [138, 157, 158, 160] have demonstrated that human gender can be classified from the gait signatures. Motivated by these existing studies, we believe there is a close relation between human age and gender from his/her gait sequences and there is noticeable difference between the gait appearances of two people with the same age but different gender. Fig. 7.1 shows the gait sequences of two persons with the same age (26 years old) but different gender. We can easily see from this figure that even they are of the same age, their appearances are different from each other due to their different genders.

Based on these observations, we attempt to make use of human gait signatures for age estimation to complement and enhance the understanding capability of visual surveillance systems. To the best of our knowledge, this is the first work that formally addresses this

Fig. 7.2: Flow chart of the proposed gait-based age estimation.

problem. To better characterize human gait appearance and enhance the robustness of the extracted gait features, we extract both Gabor magnitude (GM) and Gabor phase (GP) features of a gait sequence and perform multiple feature fusion to enhance the estimation performance. Our experimental results clearly demonstrate the feasibility of using gait signatures to estimate human age, and our proposed method can obtain good performance in age estimation.

The remaining of this chapter is organized as follows. Section 7.2 introduces the details of our proposed age estimation framework extraction. Section 7.3 discusses the enhanced age estimation by using multiple feature fusion. The experimental results and discussions are presented in Section 7.4. We conclude the paper and suggest the future work in Section 7.5.

7.2 Proposed Approach

The flow chart of our proposed gait-based human age estimation is shown in Fig. 7.2. We first extract the feature of each gait sequence, and use a label encoding algorithm to convert the age value to a binary label sequence. To mitigate the curse of dimensionality problem, we learn a multi-label-guided (MLG) subspace to reduce the dimensions of the extracted gait features. Lastly, we perform recognition by using a multi-label k -nearest neighbors (ML-KNN) classifier and decode the age information of the label vector.

7.2.1 Feature Extraction

While there have been a number of gait feature extraction techniques proposed in the literature [131–141], the gait energy image (GEI) [133] is apparently one of the most popular and successful gait representation methods owing to its superb robustness and excellent effectiveness. Hence, we use GEI for gait representation.

Fig. 7.3: Normalized and aligned silhouette images of a gait sequence. The rightmost image is the GEI.

Given a preprocessed binary gait silhouette image $B_t(x, y)$ at time t in a sequence, the gray-level GEI is defined as [133]:

$$G(x, y) = \frac{1}{N_T} \sum_{t=1}^N B_t(x, y) \quad (7.1)$$

where N_T is the number of frames in a complete cycle of a silhouette sequence, t is the frame number of the sequence, and x and y denote the image coordinates. Fig. 7.3 shows the corresponding GEI image of the gait sequence of one subject. Intuitively, a pixel with higher intensity value in GEI means that human walking occurs more frequently at this position and reflects the dynamic gait feature and that with lower intensity reflects the static gait feature.

7.2.2 Label Encoding

As mentioned earlier, existing multi-label learning techniques mainly deal with ensembles of different classes, while the age of a person can only assume a nonnegative value belonging to a range $(0, C)$, where C is usually less than 100. To make conventional multi-label learning suitable for our task, we devise an effective label encoding scheme to convert each age value into a binary sequence. We also include the gender information in the converted label vector to better exploit the dependency between the human age and gender. Fig. 7.4 illustrates the proposed label encoding approach. For each gait sample, we can obtain a multi-label vector.

7.2.3 MLG Subspace Learning

Let $X = [x_1, x_2, \dots, x_N]$ be the gait feature set and $L = [L_1, L_2, \dots, L_N]$ be the corresponding label set obtained by performing the label encoding. The objective is to learn

a low-dimensional subspace W to maximize the dependency between the feature set X and the label set L , and to address the problem of curse of dimensionality.

There are several possible criteria that can be used to measure information dependency in the literature [143–147], and we apply here a conventional criterion: Hilbert-Schmidt Independence Criterion (HSIC) [143], defined as follows:

$$HSIC(X, L, P_{XL}) = (N - 1)^{-2} \text{tr}(HKG) \quad (7.2)$$

where P_{XL} is the joint distribution of X and L , and $\text{tr}(A)$ is the trace of matrix A , $K = [K_{ij}]_{N \times N}$ and $G = [G_{ij}]_{N \times N}$ are the matrices of inner product of elements in X and L , respectively, $H = [H_{ij}]_{N \times N}$, $H_{ij} = \delta_{ij} - 1/N$, $\delta_{ij} = 1$ when $i = j$ and 0 otherwise. Since the normalization term in (7.2) doesn't affect the optimization procedure, we can drop it and optimize only $\text{tr}(HKG)$.

Let $Z = W^T X$, where W is the projection matrix to be sought, then $K = \langle Z, Z \rangle = X^T W W^T X$ and $G = L^T L$. Now, Eq. (7.2) can be rewritten as solving the following optimization problem

$$\begin{aligned} W &= \arg \max_W \text{tr}(H X^T W W^T X H G) \\ \text{subject to } & W^T W = I. \end{aligned} \quad (7.3)$$

Let $w_1, w_2, \dots, w_q$ be the eigenvectors of the following eigenfunction problem corresponding to the q largest eigenvalues ordered according to $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$,

$$H X^T G X H w = \lambda w \quad (7.4)$$

The projection matrix can be obtained as $W = [w_1, \dots, w_q]$. Interested readers are referred to [143] for details.

7.2.4 Recognition and Age Decoding

By using the proposed MLG subspace learning algorithm, the gait feature can be projected onto a low-dimensional feature space as $Z = [z_1, z_2, \dots, z_N] = W^T [x_1, x_2, \dots, x_N]$. We then apply the ML-KNN [142] classifier for recognition due to its effectiveness and efficacy. In our experiments, the neighborhood size k is empirically set to be 10.

After obtaining the label vector $L_T = [b_7^T, b_6^T, b_5^T, b_4^T, b_3^T, b_2^T, b_1^T, b_0^T]$ of a test sample T , we then decode the label vector into age information as follows:

$$A_T = b_7^T \times 2^6 + b_6^T \times 2^5 + b_5^T \times 2^4 + b_4^T \times 2^3 + b_3^T \times 2^2 + b_2^T \times 2^1 + b_1^T \quad (7.5)$$

As each gait sequence can be segmented into several walking periods, and each walking period corresponds to a GEI image, we can further fuse the recognition result of each GEI image to enhance the robustness of the age estimation.

Suppose that there are N_C walking periods in a test gait sequence and the output of these GEIs are A_t , $t = 1, 2, \dots, N_C$, we take the average of A_t 's as the final estimated age value A_f

$$A_f = \frac{1}{N_C} \sum_{t=1}^{N_C} A_t \quad (7.6)$$

where A_f is the final estimated age result of the test gait sequence.

7.3 Multiple Feature Fusion

The proposed framework employs only one single feature (GEI feature) for age estimation. Recent advances in information fusion have demonstrated that multiple feature fusion usually can achieve better performance and robustness, and outperforms single feature-based recognition method on real world applications [155]. It is therefore desirable to improve the age estimation precision by using multiple feature fusion. Gabor wavelet can extract useful discriminative information by the convoluting the original image with a set of Gabor kernels of different scales and orientations. Moreover, Gabor functions are similar to the receptive field profiles of the mammalian cortical simple cells, and it has been widely applied in many computer vision applications such as face recognition [148–151, 153] and gait recognition [152]. Motivated by these observations, we propose to make use of Gabor wavelet to represent and extract multiple gait features for age estimation. Different from conventional Gabor features which are extracted from either the Gabor magnitude (GM) [148–151, 153] or Gabor phase (GP) [152] information, we propose to use both the GM and GP features for gait representation. Moreover, most existing Gabor-based feature representation methods downsample each Gabor feature

Fig. 7.4: (a) Real parts of 40 Gabor kernels. (b) GM representation of a GEI. (c) GP representation of a GEI.

and concatenate into a new feature vector, which however may compromise some useful information for the subsequential classification. We therefore employ all of these Gabor representations for age estimation via multiple feature fusion. By doing so, we aim to retain most useful information by avoiding downsampling in the feature representation.

The GM and GP features are defined as follows

$$m_{\mu,\nu}(x, y) = \sqrt{Im^2(O_{\mu,\nu}(x, y)) + Re^2(O_{\mu,\nu}(x, y))} \quad (7.7)$$

$$p_{\mu,\nu}(x, y) = \arctan \frac{Im(O_{\mu,\nu}(x, y))}{Re(O_{\mu,\nu}(x, y))} \quad (7.8)$$

where $Re(O_{\mu,\nu}(x, y))$ and $Im(O_{\mu,\nu}(x, y))$ denote the real and imaginary parts of $O_{\mu,\nu}(x, y)$, $m_{\mu,\nu}(x, y)$ and $p_{\mu,\nu}(x, y)$ are the GM and GP features of the pixel located at (x, y) of orientation μ and scale ν , respectively.

Similar to that of [148–153], we also choose Gabor functions with eight orientations $\mu = \{0, 1, \dots, 7\}$ and five scales $\nu = \{0, 1, \dots, 4\}$, and obtain a total of 40 Gabor kernel functions. The value of each parameter follows the setting in [150]: $\sigma = 2\pi$, $k_{max} = \pi/2$, $f = \sqrt{2}$. Hence, we extract 80 Gabor representations (40 GM features and 40 GP features) from the GEI feature of a gait sequence. Fig. 7.5 shows the normalized real part of 40 kernels and the GM and GP features of a GEI.

After extracting 80 Gabor-based representation of a gait sequence, we perform the following multiple feature fusion for age estimation:

Fig. 7.5: Flow chart of multiple feature fusion-based age estimation framework.

Step 1: For each GM and GP feature f_i , $1 \leq i \leq 80$, perform single feature-based age estimation as shown in Fig. 7.6 to estimate its age value Age_i .

Step 2: Perform score-based fusion to estimate the age of a gait sequence as

$$Age = S(Age_i) \quad (7.9)$$

where $S(\cdot)$ is the fusion function. In our experiments, we apply the simplest average function to fuse them, where $S(f_i) = \frac{1}{N} \sum_{i=1}^N Age_i$. It is likely to have a more sophisticated fusion function that may further improve the estimation accuracy.

The procedure of the multiple feature fusion algorithm is illustrated in Fig. 7.6.

Table 7.1: Detailed information of all the subsets of the USF database.

Dataset	Condition	No. of subjects	Age range	Different between the gallery set
Gallery	(G, A, R, NB, M/N)	122	[19, 59]	-
Probe A	(G, A, L, NB, M/N)	122	[19, 59]	View
Probe B	(G, B, R, NB, M/N)	54	[19, 49]	Shoe
Probe C	(G, B, L, NB, M/N)	54	[19, 49]	View, Shoe
Probe D	(C, A, R, NB, M/N)	121	[19, 59]	Surface
Probe E	(C, B, R, NB, M/N)	60	[19, 49]	Shoe, Surface
Probe F	(C, A, L, NB, M/N)	121	[19, 49]	View, Surface
Probe G	(C, B, L, NB, M/N)	60	[19, 49]	View, Shoe, Surface
Probe H	(G, A, R, BF, M/N)	120	[19, 59]	Briefcase
Probe I	(G, B, R, BF, M/N)	60	[19, 49]	Briefcase, Shoe
Probe J	(G, A, L, BF, M/N)	120	[19, 59]	Briefcase, View
Probe K	(G, A/B, R, NB, M)	33	[19, 46]	Time, Clothing, Shoe
Probe L	(C, A/B, R, NB, N)	33	[19, 46]	Surface, Time, Clothing, Shoe

7.4 Experiments and Results

To verify the effectiveness of the proposed approach, we have performed extensive experiments on the well-known USF gait database [111]. We describe in this section the details of our experiments and analyze the results.

7.4.1 Data Preparation

While there are several benchmark gait databases available for gait analysis in the literature [111, 131, 132, 134], most of them don't contain age information of the gait data except the USF database. Hence, we have used the USF database to conduct age estimation experiments. The USF [3] database consists of 1870 gait sequences from 122 subjects, of whom 85 are male and 37 are female. For each subject, the variations include change in viewpoint (Left or Right), change in shoe type (A or B), change in walking surface (Grass or Concrete), change in carrying condition (Briefcase or No Briefcase), and elapsed time (May or November). There are one gallery set and 12 probe sets in the USF database. Table 7.1 describes the detailed information of each subset. Interested readers are referred to [111] for details of this database.

In our experiments, we apply the Leave-One-Person-Out (LOPO) strategy to investigate the effectiveness of the proposed age estimation framework on all the 13 subsets. In other words, the gait samples of one subject in each subset were selected for testing and the remaining were used for training. Suppose there are N_p subjects in each dataset,

we repeated this procedure N_p times and took the average as the final age estimation results.

7.4.2 Results

In our experiments, the mean absolute error (MAE) measure [85, 87–94, 96] has been used to evaluate the performance of the proposed method, which is defined as the average of the absolute errors between the estimated and the ground truth age values, defined as:

$$MAE = \sum_{i=1}^{N_t} |\hat{A}_i - A_i| / N_t \quad (7.10)$$

where $\hat{A}_i$ and A_i are the estimated and ground truth age values for the i th test sequence, and N_t is the number of the test gait sequences.

1) Performance of the Proposed Method: We first conducted experiments to calculate the performance of the proposed method. Four types of GEI representations were tested in the experiments: the first is the original gray GEI image, the second is the magnitude of Gabor filtered GEI features as suggested in [41] (downsampled and concatenated GM features), the third is the phase of Gabor filtered GEI features as suggested in [150] (downsampled and concatenated GP features), and the fourth is the proposed GM plus GP fusion method. Table 7.2 tabulates the MAEs for all the 13 sets (one gallery set and 12 probe sets) of the USF database. It’s evidenced from the results that

1) The Gabor filtered GEI feature consistently outperforms the original GEI feature for age estimation.

2) Multiple feature fusion method consistently outperforms single feature methods in terms of MAEs.

2) Comparison with Existing Human Age Estimation Methods: We then conducted experiments to compare the performance of the proposed MLG method with existing facial image-based age estimation methods. As the Quadratic Model (QM) [100] is the common benchmark age estimation algorithm and the discriminative aging manifold analysis method, called Conformal Embedding Analysis (CEA) [90], is one representative work on facial image-based age estimation, we compared our proposed method with these two existing methods [90, 100]. Moreover, we also compared two widely used classification

Table 7.2: MAEs (years old) on the USF gait database using different features.

Method	Original GEI	GM of GEI	GP of GEI	GM + GP
Gallery	6.83	6.54	6.62	6.34
Probe A	6.48	6.12	6.17	5.96
Probe B	7.38	7.08	7.12	6.84
Probe C	7.06	6.74	6.68	6.52
Probe D	7.09	6.72	6.78	6.56
Probe E	6.25	5.92	5.84	5.54
Probe F	7.67	7.32	7.28	7.05
Probe G	6.75	6.42	6.48	6.22
Probe H	6.85	6.54	6.46	6.31
Probe I	7.10	6.84	6.88	6.56
Probe J	6.61	6.24	6.18	5.96
Probe K	7.16	6.84	6.74	6.52
Probe L	6.70	6.42	6.38	6.24

algorithms including support vector machine (SVM) and nearest neighbor classifier with the proposed method in age estimation. Since a multiple feature fusion method outperforms the original GEI feature for gait-based age estimation as demonstrated in the previous experiments, we applied multiple feature fusion approach (both GM and GP as the features of gait sequences) for age estimation in the following experiments. Table. 7.3 shows the MAEs of the five methods on all the 13 sets of the USF database. It is evidenced from the results that the proposed method consistently outperforms the other four compared methods for gait-based age estimation.

We provide a reason why the existing methods do not perform as well as the proposed MLG method. As mentioned in Section 7.1, manifold analysis based age estimation methods usually need lots of facial images of different ages to model the aging manifold. In real-world visual surveillance applications, it’s difficult to collect a large number of gait sequences of a person of different ages. For example, in this study, the largest gait database is the gallery and probe A datasets, containing 122 subjects corresponding to 656 and 663 gait samples; hence, the aging manifold learned may not be accurate. To further illustrate this point, we applied the CEA method to learn the aging manifolds of the gallery and probe A datasets and plot them in Fig. 7.7. It’s easy to see that the learned aging manifold doesn’t appear as a dispersal type nor follow a curved chronological

Table 7.3: MAEs (years old) on all the 13 sets of the USF database using different methods.

Method	QM	CEA	SVM	NN	MLG
Gallery	6.74	6.64	7.12	7.38	6.34
Probe A	7.32	6.56	6.66	6.47	5.96
Probe B	7.30	7.12	7.52	7.48	6.84
Probe C	7.08	6.94	8.12	8.35	6.52
Probe D	6.75	6.62	6.98	6.93	6.56
Probe E	6.04	6.32	6.52	6.56	5.54
Probe F	7.31	7.12	7.24	7.19	7.05
Probe G	7.11	6.71	6.92	6.84	6.22
Probe H	6.72	6.42	6.76	6.83	6.31
Probe I	7.47	6.82	6.91	6.94	6.56
Probe J	6.24	6.32	6.24	6.34	5.96
Probe K	6.81	6.77	6.74	6.77	6.52
Probe L	6.54	6.44	6.58	6.61	6.24

trend. Please note that this observation was also made in another facial age estimation work [156], where manifold learning algorithms failed to discover an apparent aging trend when the number of data is not large enough. We also present the aging manifolds learned by PCA on these two dataset, and it is easy to see while the aging trends were not discovered by these two methods, the aging manifold learned by PCA and CEA are quite different.

Second, QM is one regression-based age estimation method and the quadratic function is used to estimate the human age. However, even it has achieved reasonably good performance, there isn't a strong justification that the relation between the human facial images (or gait sequences) and the ages closely follows a quadratic model. However, the proposed MLG method doesn't need such an assumption and hence performs better than the QM method for age estimation.

Lastly, SVM and NN are two commonly used classification algorithms and they predict the age of the test sample by using only the age information and ignore the gender information of the training samples. Different from them, our proposed MLG has learned a feature subspace by utilizing both gender and age information for age estimation. Hence, the age and gender relation could be well exploited in the reduced feature space and better age estimation performance can be obtained.

Fig. 7.6: Aging manifold learned by the (a) PCA method on the Gallery set, (b) CEA method on the Gallery set, (c) PCA method on the Probe A set and (d) CEA method on the Probe A set, respectively. The color data points of age from 19 to 59 are colored from blue to red.

3) Comparison with the Existing Age Estimation with Gender Classification Method:

Previous work has shown that the performance of an age estimation system was affected by human gender information, and better performance can be obtained when perform age estimation on different gender groups [15]. Specifically, for a given test sample, we can recognize the gender of a face and then perform age estimation corresponding to that gender. Based on this observation, we compare here our method with the existing age estimation with gender classification method. We have applied the QM and CEA methods for age estimation in two different manners: 1) age estimation without gender

Fig. 7.7: Age range distribution of the selected USF subset database.

classification (as the approach used in the previous experiments), and 2) gender classification (GC) + age estimation, first recognizing the gender of a test gait sample and then performing age estimation corresponding to that gender. As the Gallery and Probe A sets contain most subjects among all the 13 sets of the USF database and the original datasets are biased in terms of age and gender distribution (only 37 subjects are female), we selected two subsets from the USF gallery and probe A datasets, respectively, containing the same subjects under the age range of $[15, 39]$ so that the age distribution is more uniform, and 79 subjects were selected, in which 45 are male and 34 are females. Fig. 7.8 shows the age distribution of the two selected subsets.

We applied the LOPO strategy to perform age estimation on the two selected subsets. The gait samples of each subject were selected for testing and the remaining were used for training, and we repeated this procedure 79 times and took the average as the final age estimation results. Table 7.4 shows the MAEs obtained by using different age estimation methods on the two subsets of the USF database. It is evidenced from the results that performing age estimation with gender classification can improve the age estimation performance as compared to that without gender classification.

While improved performance can be obtained when gender classification is performed before age estimation, this two-step approach needs to know all the gender information of the training samples, and samples without gender information cannot be utilized for

Table 7.4: MAEs (years old) on the two subsets of the USF database using different methods.

Method	QM	CEA	GC+QM	GC+CEA	MLG
Gallery	5.82	5.74	5.56	5.44	5.42
Probe A	5.94	5.87	5.72	5.59	5.56

(a) Results on the Gallery subset.

(b) Results on the probe A subset.

Fig. 7.8: MAEs of estimation for three different methods when different percentages of training samples have missing gender labels on the gallery and probe A subsets of the USF database.

training in the gender classification phase. However, the proposed MLG method can mitigate this shortcoming, as detailed below. We randomly selected θ percentages ($0 \leq \theta \leq 50$) of gait samples from the training set and assumed that the gender information of these samples were unknown. In such scenarios, we randomly assigned 0 or 1 to the gender bit and performed multi-label learning for age estimation. Fig. 7.9 shows the MAEs of the gallery and probe A subsets versus the parameter θ , which validates the performance of the proposed method. We can see from this figure that the performance of the proposed method degrades gracefully when the percentage of training samples with missing gender information increases. We can also see from this figure that the performance of the proposed method significantly outperforms the QM and CEA methods for age estimation under this scenario, and even for the case that none of the gender information was used in the estimation.

It is interesting to understand why the proposed MLG method can still outperform

Table 7.5: MAEs (years old) on the two subsets of the USF database using different gait features.

Method	GEI	Method in [140]
Gallery	5.42	5.28
Probe A	5.56	5.40

the QM and CEA methods when the gender information is missing. As mentioned in the label encoding subsection, the gender information is only encoded as one bit in the label vector (b_0) and there are totally 8 bits in the label vector. Hence, even if the gender information is not available, it may not affect substantially the relationship of each label vector as the remaining 7 bits of age information (b_7-b_1) can still be used to exploit the correlation of the label vectors.

4) *Effects of Enhanced Gait Feature Representation:* While GEI has attained reasonably good performance in gait recognition, existing GEI-based feature representation methods treat GEI as a two-dimensional (2D) image and neglect the intrinsic dynamic characteristics of gait patterns which contain more dynamic information of the walking persons. Based on this observation, Bashir *et al.* [140] proposed a feature selection method for GEI-based gait feature representation which is less susceptible to covariate structure. In this subsection, we applied this improved GEI for gait feature representation and performed age estimation on the Gallery and Probe A subsets by using our proposed MLG method. Table 7.5 shows the MAEs obtained when different feature representation methods were used on the two subsets of the USF database. It is evidenced from the results that the enhanced GEI feature representation method can further improve the age estimation performance, which further suggests that dynamic information of the GEI features contributes positively to human age estimation.

5) *Cross-Dataset Age Estimation:* It may be interesting to see our age estimation performance on other gait sequences. As we mentioned before, while there are several benchmark gait datasets used for human gait analysis and recognition in the literature, most of them contain only a few subjects and there are only three publicly available gait databases containing more than 100 subjects used for gait recognition, namely, the SOTON, USF and CASIA databases. These databases were collected by University of Southampton, University of South Florida and Institute of Automation, Chinese

Fig. 7.9: Sample frames of one subject from our gait database.

Table 7.6: MAEs (years old) on our dataset using different gait features.

Feature	MAE
Original GEI	3.86
GM of GEI	3.45
GP of GEI	3.34
GM + GP fusion	3.02

Academy of Sciences, respectively. While the SOTON and CASIA databases can enable us to perform multiview gait analysis, these two databases don't contain age information of the gait data and could not be used for age estimation. To investigate the age estimation performance of the proposed approach, we collected a small gait datasets containing 50 subjects with the age range of $[20, 30]$, of whom 30 are male and 20 are female. Each subject was asked to walk naturally along a straight line in an outdoor environment 2 times. Hence, the dataset contains a total of $50 \times 2 = 100$ gait sequences. All the videos have the same resolution of 320×240 and some sample frames are shown in Fig. 7.10.

Given each gait sequence, the human silhouettes can be easily extracted by background subtraction and thresholding in a way similar to the approach proposed in our work described in [159]. We learned a MLG subspace by using the Gallery subset and performed age estimation on our dataset. As there are two gait sequences collected for each subject, we took the average as the final age estimation. Table 7.6 lists the MAEs obtained from our dataset. This result shows the robustness of our proposed age estimation approach to human age estimation.

7.4.3 Discussion

From the above experimental results, we have the following observations:

- 1) Gait signature can be used as an effective cue for human age estimation at a distance, which can enhance the understanding capability of many existing visual surveil-

lance systems.

2) The Gabor filtered GEI feature outperforms the original GEI feature for gait-based human age estimation. This is probably due to that Gabor functions are similar to the receptive field profiles of the mammalian cortical simple cells, which have demonstrated a better gait representation capability.

3) The multiple feature fusion method consistently outperforms single feature methods for gait-based age estimation, which implies that multiple features can provide complementary information for feature representation, resulting in a better discriminative power for estimation/classification.

4) The proposed age estimation method can also obtain good performance on cross-dataset age estimation experiments, which shows the robustness of our approach to gait-based human age estimation.

7.5 Chapter Summary

We have studied the problem of gait-based human age estimation at a distance for visual surveillance. To our knowledge, the problem has not been formally addressed in the literature. Since gait appearances are different for males and females even under the same age group, we have devised a multi-label-guided (MLG) subspace to better characterize the correlation information between gender of age information. Moreover, as the age of a person can only assume a nonnegative value and existing multi-label learning techniques can handle only binary label sequences, we have also proposed an effective label encoding scheme to convert each age value to a binary sequence, making conventional multi-label learning suitable for our task. To better characterize human gait appearance, we extract a set of over-complete Gabor features including both Gabor magnitude (GM) and Gabor phase (GP) information of a gait sequence and perform multiple feature fusion to enhance the estimation performance. Experimental results have demonstrated the efficacy of the proposed approach. Partial work presented in this chapter was published in [168] and [169].

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this thesis, we have presented several enhanced linear subspace learning algorithms for face and gait analysis. The main contributions of this thesis are to explore, analyze and propose solutions to the problems of feature extraction in many pattern recognition and computer vision applications such as face recognition. These proposed algorithms are experimentally evaluated and compared with some existing popular and state-of-the-art methods.

We first investigated the different importances of different samples and feature elements in the training sets. We proposed a doubly weighted subspace learning approach by imposing larger weights on samples which have low similarity to others in the same class and applying different emphases on different feature elements according to their importances for recognition and human age estimation. We constructed two weighting matrices based on pairwise similarity of samples within a same class and discriminant score of each pixel within a sample to duly emphasize the between-sample and within-sample features. We then incorporated these two weighting matrices into three conventional subspace learning methods, namely PCA, LDA, and NMF, to devise the corresponding weighted methods. Experimental results on the CMU PIE and Extended Yale B face databases were presented to demonstrate the efficacy of the proposed approach.

Second, we investigated the SSS problem of some newly proposed manifold-based linear subspace algorithms. To address this problem, we proposed a parametric regularization approach to regulate the eigenfeature space and perform feature extraction in a

whole feature space rather than a reduced feature subspace of PCA. We applied this approach into four newly proposed subspace learning methods, namely LPP, DLPP, OLPP and ODLPP, to devise the corresponding regularized methods. Experimental results on the ORL and FERET face databases were presented to demonstrate the efficacy of the proposed approach.

Third, we studied subspace learning in a cost-sensitive setting. To our knowledge, this is the first work that formally addresses the cost-sensitive problem in subspace learning. Our approach incorporates a cost matrix, which specifies the different costs associated with misclassifications of subjects, into three popular subspace learning algorithms, namely PCA, LDA and LPP, and devise the corresponding cost-sensitive methods, to achieve a minimum overall recognition loss by performing recognition in the low-dimensional subspaces derived. Experimental results on the AR and FERET face databases were presented to demonstrate the efficacy of the proposed approach.

Then, we investigated subspace learning with limited number of training samples. Since existing subspace learning algorithms degrade heavily when the number of training samples per class is small because such limited samples are usually not able to effectively characterize the geometrical structure of the original sample space for learning an effective feature subspace, we proposed using the NFL and NFS metrics rather than the conventional NN metric to seek discriminative feature spaces for feature extraction. Experimental results on face recognition and facial expression recognition were presented to demonstrate the efficacy of the proposed methods.

Lastly, we investigated the problem of estimating human ages from gait signatures. To our knowledge, this problem has not been formally addressed in the literature. As human age assumes only nonnegative value and existing multi-label learning techniques mainly deal with ensembles of binary classes, we devised an effective label encoding scheme to convert each age value to a binary sequence, making conventional multi-label learning suitable for our task. To better characterize human gait appearance and enhance the robustness of the proposed age estimation method, we extracted a set of over-complete Gabor features including both Gabor magnitude (GM) and Gabor phase (GP) information of a gait sequence and perform multiple feature fusion to enhance the age estimation performance. Our experimental results clearly demonstrated the feasibility of using gait signatures to estimate human age and the efficacy of the proposed method.

8.2 Future Work

In this section, we present some research directions which can be explored in the future.

For our doubly weighted subspace learning approach, we have only defined a very simple function to compute the representative power of each sample, which is somehow empirical rather than theoretical. How to design a more elaborate representative function in theory to better characterize the within-sample weighting information of samples to further improve the performance remains unsolved. For the within-sample weighting, we have only computed the discriminative score of each pixel and employed such discriminative score as within-sample weighting. We suggest to design a subject-specific within-sample weighting method, in which the different importances of different feature elements for subspace learning could be better utilized.

For the parametric regularized subspace learning approach, we have only used a parametric regularization $1/f^a$ ($a > 0$) to perform feature regularization. How to investigate other function forms which could better regulate the eigenfeature space is an interesting research direction of our future work.

For the cost-sensitive subspace learning approach, the cost matrix was pre-specified by user. In many real world applications, it is not easy for the users to provide a clear cost for the system specifically. How to learn a cost matrix for cost-sensitive subspace learning is another interesting research problem. One potential solution is to use a cost range rather than a specific cost for learning a cost-sensitive feature subspace such that it would be much more convenience to determine a cost range rather than a specific cost for real world applications.

For the proposed NFL- and NFS-based subspace learning algorithms, we have used feature lines and feature spaces to generate more virtual samples to enhance the learned feature space. However, there may also be some noise points because feature lines and feature spaces are naturally linear and many real world data, such as face samples, generally distributed on a nonlinear manifold. Hence, it is desirable to apply a more elaborate metric to generate virtual samples to handle the nonlinear distributions for subspace learning.

For the multi-label subspace learning method, we implicitly assume the importance of each bit in the the coded age and gender label vector is the same. Since misclassification

of the label in different bits lead to different age errors, it is more acceptable to assign different weights on different bits to better learn a multi-label subspace for gait feature extraction. Moreover, we have only fused the GM and GP feature by using a simple average function, and it is likely to have a more sophisticated fusion function that may further improve the estimation accuracy.

Author's Publications

- (1) **Jiwen Lu** and Yap-Peng Tan, "A doubly weighted approach for appearance-based subspace learning methods," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 5, no. 1, pp. 71-81, 2010.
- (2) **Jiwen Lu** and Yap-Peng Tan, "Regularized locality preserving projections and its extensions for face recognition," *IEEE Transactions on Systems, Mans, and Cybernetics, Part B (TSMC-B)*, vol. 40, no. 3, pp. 958-963, 2010.
- (3) **Jiwen Lu** and Yap-Peng Tan, "Gait-based human age estimation," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 5, no. 4, pp. 761-770, 2010.
- (4) **Jiwen Lu** and Yap-Peng Tan, "Nearest feature space analysis for classification," *IEEE Signal Processing Letters (SPL)*, vol. 18, no. 1, pp: 55-58, 2011.
- (5) **Jiwen Lu** and Yap-Peng Tan, "Uncorrelated discriminant nearest feature line analysis for face recognition," *IEEE Signal Processing Letters (SPL)*, vol. 17, no. 2, pp. 185-188, 2010.
- (6) **Jiwen Lu** and Yap-Peng Tan, "Uncorrelated discriminant simplex analysis for view-invariant gait signal computing," *Pattern Recognition Letters (PRL)*, vol. 31, no. 5, pp. 382-393, 2010.
- (7) **Jiwen Lu** and Yap-Peng Tan, "Improved discriminant locality preserving projections for face and palmprint recognition," *Neurocomputing*, 2011, to appear.
- (8) Nini Liu, **Jiwen Lu** and Yap-Peng Tan, "Joint subspace learning for view-invariant gait recognition," *IEEE Signal Processing Letters (SPL)*, vol. 18, no. 7, pp. 431-434, 2011.

- (9) **Jiwen Lu**, Yap-Peng Tan, and Gang Wang, “Discriminative multi-manifold analysis for face recognition from a single training sample per person,” *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011, to appear.
- (10) **Jiwen Lu** and Yap-Peng Tan, “Cost-sensitive subspace learning for face recognition,” *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2661-2666, 2010.
- (11) **Jiwen Lu** and Yap-Peng Tan, “Ordinary preserving manifold analysis for human age estimation,” *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 90-95, 2010. (oral presentation)
- (12) **Jiwen Lu** and Yap-Peng Tan, “Fusing shape and texture information for facial age estimation,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.
- (13) **Jiwen Lu** and Yap-Peng Tan, “Locality repulsion projections for image-to-set face recognition,” *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2011. (oral presentation)
- (14) **Jiwen Lu** and Yap-Peng Tan, “Adaptive maximum margin criterion for image classification,” *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2011. (oral presentation, best paper candidate)
- (15) **Jiwen Lu** and Yap-Peng Tan, “Cost-sensitive subspace learning for human age estimation,” *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 1593-1596, 2010. (oral presentation)
- (16) **Jiwen Lu** and Yap-Peng Tan, “View recognition of human gait sequences in videos,” *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 2457-2460, 2010. (oral presentation)
- (17) **Jiwen Lu** and Yap-Peng Tan, “Gait-based human age estimation,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1718-1721, 2010. (oral presentation)

- (18) **Jiwen Lu** and Yap-Peng Tan, “Doubly weighted nonnegative matrix factorization for imbalanced face recognition,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 877-880, 2009.
- (19) **Jiwen Lu** and Yap-Peng Tan, “Two-directional two-dimensional discriminant locality preserving projections for image recognition,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1753-1756, 2009.
- (20) **Jiwen Lu** and Yap-Peng Tan, “Uncorrelated multilinear geometry preserving projections for multimodal biometrics recognition,” *Proceedings of IEEE International Symposium on Circuit and Systems (ISCAS)*, pp. 2601-2604, 2009. (oral presentation)
- (21) **Jiwen Lu** and Yap-Peng Tan, “Enhanced face recognition using tensor neighborhood preserving discriminant projections,” *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 1916-1919, 2008.
- (22) Nini Liu, **Jiwen Lu** and Yap-Peng Tan, “Set-to-set gait recognition across varying views and walking conditions,” *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2011. (oral presentation)
- (23) Tian Xia, **Jiwen Lu**, and Yap-Peng Tan, “Face recognition using an enhanced age simulation method,” *Proceedings of IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2011, to appear, (oral presentation)
- (24) Meng Cui, **Jiwen Lu**, and Yap-Peng Tan, “A comparative study of age-invariant face recognition with different feature representations,” *Proceedings of International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2010.
- (25) Nini Liu, **Jiwen Lu**, Yap-Peng Tan, and Zhenzhong Chen, “Enhanced gait recognition based on weighted dynamic feature,” *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 3581-3584, 2009.

Bibliography

- [1] M. Turk, and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.
- [2] P. N. Belhumeur, J. P. Hefanah, 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.
- [3] A. M. Martinez, and A. C. Kak, “PCA versus LDA,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, 2001.
- [4] M.-H. Yang, “Kernel eigenfaces versus kernel Fisherfaces: face recognition using kernel methods,” *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 215-220, 2002.
- [5] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.
- [6] S. T. Roweis, and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000.
- [7] M. Belkin, and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Computation*, vol. 15, no. 6, pp. 1373-1396, 2003.
- [8] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, “Face recognition using Laplacianfaces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 3, pp. 328-340, 2005.

- [9] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1208-1213, 2005.
- [10] X. He, D. Cai, and J. Han, "Learning a maximum margin subspace for image retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 189-201, 2008.
- [11] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal Laplacianfaces for face recognition," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3608-3614, 2006.
- [12] D. Cai, X. He, K. Zhou, J. W. Han, and H. J. Bao, "Locality sensitive discriminant analysis," *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 708-713, 2007.
- [13] D. Cai, X. He, Y. Hu, J. Han, and T. Huang, "Learning a spatially smooth subspace for face recognition," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.
- [14] D. Cai, X. He, and J. Han, "Spectral regression for efficient regularized subspace learning," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1-8, 2007.
- [15] S. Yan, J. Liu, X. Tang, and T. Huang, "A parameter-free framework for general supervised subspace learning," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 1, pp. 69-75, 2007.
- [16] J. Yang, S. Yan, Y. Fu, X. Li, and T. Huang, "Non-negative graph embedding," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.
- [17] S. Yan, D. X, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, 2007.

- [18] Y. Fu, S. Yan, and T. S. Huang, "Classification and feature extraction by simplex-ization," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp: 91-100, 2008.
- [19] Y. Fu, S. Yan, and T. Huang, "Correlation metric for generalized feature extraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2229-2235, 2008.
- [20] W. Yu, X. Teng, and C. Liu, "Face recognition using discriminant locality preserving projections," *Image and Vision Computing*, vol. 24, no. 3, pp. 239-248, 2006.
- [21] L. Zhu, and S. Zhu, "Face recognition based on orthogonal discriminant locality preserving projections," *Neurocomputing*, vol. 70, no. 7-9, pp. 1543-1546, 2007.
- [22] X. Yu, and X. Wang, "Uncorrelated discriminant locality preserving projections," *IEEE Signal Processing Letters*, vol. 15, no. 3, pp. 383-394, 2008.
- [23] E. Kokiopoulou, and Y. Saad, "Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2143-2156, 2007.
- [24] J. Yang, D. Zhang, J.-Y. Yang, B. Niu, "Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 650-664, 2007.
- [25] H. Wang, S. Chen, Z. Hu, and W. Zheng, "Locality-preserving maximum information projecton," *IEEE Transactions on Neural Networks*, vol. 19, no. 4, pp. 571-585, 2008.
- [26] H. Hu. "Orthogonal neighborhood preserving discriminant analysis for face recognitions," *Pattern Recognition*, vol. 41, no. 6, pp. 2045-2054, 2008.
- [27] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variant," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 846-853, 2005.

- [28] J. Yang, D. Zhang, A. F. Frangi, and J. Y. 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.
- [29] J. Ye, R. Janardan, and Q. Li. "Two-dimensional linear discriminant analysis," *Proceedings of Advances in Neural Information Processing Systems*, pp. 1569-1576, 2004.
- [30] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," *Proceedings of Advances in Neural Information Processing Systems*, pp. 499-506, 2005.
- [31] J. Lu, and Y.-P. Tan, "Two-directional two-dimensional discriminant locality preserving projections for image recognition," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1753-1756, 2009.
- [32] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: multilinear principal component analysis of tensor objects," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 18-39, 2008.
- [33] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. J. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 212-220, 2007.
- [34] W. Zuo, D. Zhang, J. Yang and K. Wang, "BDPCA plus LDA: A novel fast feature extraction technique for face recognition," *IEEE Transactions on Systems, Mans, and Cybernetics, Part B*, vol. 36, no. 4, pp. 946-953, 2006.
- [35] N. K. Logothetis, and D. L. Sheinberg, "Visual object recognition," *Annual Review of Neuroscience*, vol. 19, pp. 509-522, 1996.
- [36] D. D. Lee, and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.
- [37] D. D. Lee, and S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing System*, vol. 14, pp. 556-562, 2001.

- [38] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with applications to frontal face verification," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 683-695, 2006.
- [39] I. Kotsia, S. Zafeiriou, and I. Pitas, "A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 588-595, 2007.
- [40] Y. Xue, C. S. Tong, W.-S. Chen, W. Zhang, and Z. He, "A modified non-negative matrix factorization algorithm for face recognition," *Proceedings of IEEE International Conference on Pattern Recognition*, 495-498, 2006.
- [41] S. Z. Li, X. Hou, H. J. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 207-212, 2001.
- [42] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.
- [43] P. M. Alberto, J. M. Carazo, K. Kochi, and D. P. M. Roberto, "Nonsmooth non-negative matrix factorization (nsNMF)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 403-415, 2006.
- [44] T. Zhang, B. Fang, Y. Y. Tang, G. He, and J. Wen, "Topology preserving non-negative matrix factorization for face recognition," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 574-584, 2008.
- [45] J. Yang, S. Yan, Y. Fu, X. Li, and T. Huang, "Non-negative graph embedding," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.
- [46] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," *Proceedings of IEEE International Conference on Data Mining*, pp. 63-72, 2008.

- [47] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space," *IEEE Transactions on Neural Networks*, vol. 19, no. 6, pp. 1090-1100, 2008.
- [48] I. Buciu, "Nonnegative matrix factorization, a new tool for feature extraction theory and applications," *International Journal of Computers, Communications and Control*, vol. 3, pp. 67-74, 2008.
- [49] X. Xie, and K.-M. Lam, "Gabor-based kernel PCA with doubly nonlinear mapping for face recognition with a single face image," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2481-2492, 2006.
- [50] K.-H. Lin, K.-M. Lam, and W.-C. Siu, "Spatially eigen-weighted Hausdorff distances for human face recognition," *Pattern Recognition*, vol. 15, no. 9, pp. 2481-2492, 2003.
- [51] T. Tan, and H. Yan, "Face recognition using the weighted fractal neighbor distance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 35, no. 4, pp. 576-582, 2005.
- [52] J. Zhou, Q. Ji, and G. Nagy, "A comparative study of local matching approach for face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2617-2628, 2007.
- [53] D. Guillaumet, J. Vitria, B. Schiele, "Introducing a weighted non-negative matrix factorization for image classification," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2447-2454, 2004.
- [54] D. Guillaumet, M. Bressan, and J. Vitria, "A weighted non-negative matrix factorization for local representations," *Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 942-947, 2001.
- [55] G. W. Stewart, "Matrix Algorithms Volumn II: Eigensystems," *Philadelphia, PA: SIAM*, 2001.

- [56] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp: 165-175, 1989.
- [57] D. Q. Dai, and P. C. Yuen, "Face recognition by regularized discriminant analysis," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 37, no. 4, pp. 1080-1085, 2007.
- [58] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383-394, 2008.
- [59] J. Lu, K. N. Plataniotis, and V. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181-191, 2005.
- [60] B. Moghaddam, and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696-710, 1997.
- [61] F. Song, D. Zhang, D. Mei, and Z. Guo, "A multiple maximum scatter difference discriminant criterion for facial feature extraction," *IEEE Transactions on Systems, Mans, and Cybernetics, Part B*, vol. 37, no. 6, pp. 1599-1606, 2007.
- [62] Y.-F. Li, J. T. Kwok, and Z.-H. Zhou, "Cost-sensitive semi-supervised support vector machine," *Proceedings of National Conference on Artificial Intelligence*, Georgia, USA, July, 2010.
- [63] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358-3378, 2007.
- [64] H. Masnadi-Shirazi, and N. Vasconcelos, "Cost-sensitive boosting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI:10.1109/TPAMI.2010.71, 2010, to appear.

- [65] C. Elkan, “The foundations of cost-sensitive learning,” *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 973-978, 2001.
- [66] S. Raudys, and A. Raudys, “Pairwise costs in multiclass perceptions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1324-1328, 2010.
- [67] Z.-H. Zhou, and X.-Y. Liu, “On multi-class cost-sensitive learning,” *Proceedings of National Conference on Artificial Intelligence*, pp. 567-572, 2006.
- [68] Y. Zhang and Z.-H. Zhou, “Cost-sensitive face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1758-1769, 2010.
- [69] A. Jain, A. Ross, and S. Prabhakar, “An introduction to biometric recognition,” *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 14, no. 1, pp. 801-822, 2004.
- [70] S. Li, and J. Lu, “Face recognition using the nearest feature line method,” *IEEE Transactions on Neural Networks*, vol. 10, no. 2, pp. 439-443, 1999.
- [71] S. Li, “Content-based classification and retrieval of audio using the nearest feature line method,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 619-625, 2000.
- [72] S. Li, K. Chan, and C. Wang, “Performance evaluation of the nearest feature line method in image classification and retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1335-1339, 2000.
- [73] K. Chen, T.-Y. Wu, and H.-J. Zhang, “On the use of nearest feature line for speaker identification,” *Pattern Recognition Letters*, vol. 23, no. 14, pp. 1735-1746, 2002.
- [74] H. Wang, and L. Zhang, “Linear generalization probe samples for face recognition,” *Pattern Recognition Letters*, vol. 25, no. 8, pp. 829-840, 2004.
- [75] Y.-S. Ryu, and S.-Y. Oh, “Simple hybrid classifier for face recognition with adaptively generated virtual data,” *Pattern Recognition Letters*, vol. 23, no. 7, pp. 833-841, 2002.

- [76] O.-A. Mauricio, R. P. W. Duin, and G.-D. German, "A generalization of dissimilarity representations using feature lines and feature planes," *Pattern Recognition Letters*, vol. 30, no. 3, pp. 242-254, 2009.
- [77] W. Zheng, L. Zhao, and C. Zou, "Locally nearest neighbor classifiers for pattern classification," *Pattern Recognition*, vol. 37, no. 6, pp. 1307-1309, 2007.
- [78] Q.-B. Gao, and Z.-Z. Wang, "Center-based nearest neighbor classifier," *Pattern Recognition*, vol. 40, no. 1, pp. 346-349, 2007.
- [79] H. Du, and Q. Chen, "Rectified nearest feature line segment for pattern classification," *Pattern Recognition*, vol. 40, no. 5, pp. 1486-1497, 2007.
- [80] Y.-J. Zheng, J.-Y. Yang, J. Yang, X.-J. Wu, and Z. Jin, "Nearest neighbour line nonparametric discriminant analysis for feature extraction," *Electronics Letters*, vol. 42, no. 12, pp. 679 - 680, 2006.
- [81] Y. Pang, Y. Yuan, and X. Li, "Generalized nearest feature line for feature extraction," *Electronics Letters*, vol. 43, no. 20, pp. 1079-1080, 2007.
- [82] Y. Pang, Y. Yuan, and X. Li, "Iterative subspace analysis based on feature line distance," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 903-907, 2009.
- [83] Z. Jin, J. Yang, Z. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognition*, vol. 34, no. 7, pp. 1405-1416, 2001.
- [84] J.-T. Chien, and C.-C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1644-1649, 2002.
- [85] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang, "A study on automatic age estimation using a large database," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1-6, 2009.

- [86] G. Guo, G. Mu, Y. Fu, and T. Huang, "Human age estimation using bio-inspired features," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 112-119, 2009.
- [87] Y. H. Kwon, and N. V. Loba, "Age classification from facial images," *Computer Vision and Image Understanding*, vol. 74, no. 1, pp. 1-21, 1999.
- [88] G. Guo, Y. Fu, T. Huang, and C. Dyer, "Locally adjusted robust regression for human age estimation," *Proceedings of IEEE Workshop on Application of Computer Vision*, pp. 1-6. 2008.
- [89] G. Guo, Y. Fu, C. R. Dyer, T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178-1188, 2008.
- [90] Y. Fu, and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 578-584, 2008.
- [91] Y. Fu, Y. Xu, and T. Huang, "Estimating human age by manifold analysis of face pictures and regression on aging features," *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 1383-1386, 2007.
- [92] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," *Proceedings of ACM International Conference on Multimedia*, pp. 307-316, 2006.
- [93] X. Geng, Z. H. Zhou, and S. M. Kate, "Automatic age estimation based on facial aging patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2234-2240, 2007.
- [94] X. Geng, S.-M., Kate, Z.-H., Zhou, and L. Wang, "Facial age estimation by non-linear aging pattern subspace," *Proceedings of ACM International Conference on Multimedia*, pp. 721-724, 2008.

- [95] Z. Yang, and H. Ai, "Demographic classification with local binary patterns," *Proceedings of IEEE International Conference on Biometrics*, pp. 464-473, 2007.
- [96] J. Suo, T. Wu, S. Zhu, S. Shan, X. Chen, and W. Gao, "Design sparse features for age estimation using hierarchical face model," *Proceedings of IEEE International Conference on Face and Gesture Recognition*, pp. 1-6, 2008.
- [97] S. Yan, H. Wang, X. Tang, J. Liu, and T. Huang, "Regression from uncertain labels and its applications to soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 698-708, 2008.
- [98] S. Yan, H. Wang, X. Tang, and T. Huang, "Learning auto-structured regressor from uncertain nonnegative labels. *Proceedings of IEEE International Conference on Computer Vision*, pp. 1-8, 2007.
- [99] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. Huang, "Regression from patch-kernel," *Proceedings of IEEE International Conference on Computer Vision and Pattern recognition*, pp. 1-8, 2007.
- [100] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Towards automatic simulation of aging effects on face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 442-455, 2002.
- [101] A. Lanitis, "Comparative evaluation of automatic age progression methodologies," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 239480, 2008.
- [102] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Transactions on Systems, Mans, and Cybenetics-Part B: Cybenetics*, vol. 34, no. 2, pp. 621-628, 2004.
- [103] Y. Fu, G. Guo and T.S. Huang, "Age synthesis and estimation via faces: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1955-1976, 2010.

- [104] A. S. Georghiades, P. N. Belhumeur and 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.
- [105] K.-C. Lee, J. Ho and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol, 27, no. 5, pp. 684-698, 2005.
- [106] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illuminant, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 9, pp. 1615-1618, 2003.
- [107] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face-recognition problem," *Image and Vision Computing*, vol. 16, no. 5, pp. 295-306, 1998.
- [108] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090-1104, 2000.
- [109] A. M. Martinez and R. Benavente, "The AR face database," *CVC Technical Report*, no. 24, 1998.
- [110] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Proceedings of IEEE International Conference on Face and Gesture Recognition*, pp. 46-53, 2000.
- [111] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: data sets, performance, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162-177, 2005.
- [112] F. Wang, J. Wang, C. Zhang, and J. Kwok, "Face recognition using spectral features," *Pattern Recognition*, vol. 40, no. 10, pp. 2786-2797, 2007.

- [113] J. Yu, Q. Tian, T. Rui, and T. S. Huang, "Integrating discriminant and descriptive information for dimension reduction and classification," *IEEE Transactions on Circuit and System for Video and Technology*, vol. 17, no. 3, pp. 372-377, 2007.
- [114] Q. Liu, X. Tang, H. Liu, and S. Ma, "Face recognition using kernel scatter-difference-based discriminant analysis," *IEEE Transactions on Circuit and System for Video and Technology*, vol. 17, no. 7, pp. 1081-1085, 2006.
- [115] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: a complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230-244, 2005.
- [116] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*, , vol. 17, no. 1, pp. 157-165, 2006.
- [117] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1222-1228, 2004.
- [118] X. Wang and X. Tang, "Random sampling for subspace face recognition," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 91-104, 2006.
- [119] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowske, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450-1464, 2002.
- [120] K. C. Kwak, and W. Pedrycz, "Face recognition using an enhanced independent component analysis approach," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 530-541, 2007.
- [121] S. Yang, S. Yan, C. Zhang, and X. Tang, "Bilinear analysis for kernel selection and nonlinear feature extraction," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1442-1452, 2007.

- [122] Q. Liu, H. Liu, and S Ma, “Improving kernel fisher discriminant analysis for face recognition,” *IEEE Transactions on Circuit and Systems for video Technology*, vol. 14, no. 1, pp. 42-490, 2004.
- [123] 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, 2003.
- [124] 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, no. 1, pp. 4-13, 2005.
- [125] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face description with local binary patterns: application to face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037-2041, 2006.
- [126] M. A. X. Vasilescu, and D. Terzopoulos, “Multilinear (Tensor) image synthesis, analysis, and recognition,” *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 118-123, 2007.
- [127] Y. Fu and T. S. Huang, “Image classification using correlation tensor analysis,” *IEEE Transactions on Image Processing*, vol. 30, no. 2, pp. 226-234, 2008.
- [128] D. Xu, S. Yan, L. Zhang, H. J. Zhang, and T. S. Huang, “Reconstruction and recognition of tensor-based objects with concurrent subspace analysis,” *IEEE Transactions on Circuit and System for Video Technology*, vol. 18, no. 1, pp. 36-47, 2008.
- [129] R. Costantini, L. Sbaiz, and S. Sússtrunk, “Higher order SVD analysis for Dynamic texture synthesis,” *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 42-52, 2008.
- [130] S. H. Lee, and S. Choi, “Two-dimensional canonical correlation analysis,” *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 735-738, 2008.
- [131] M. S. Nixon, and J. N. Carter, “Automatic recognition by gait,” *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2013-2024, 2006.

- [132] N. V. Boulgouris, D. Hatzinakos, and K. N. Plataniotis, "Gait recognition: a challenging signal processing technology for biometric identification," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 78-90, 2005.
- [133] J. Han, and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316-322, 2006.
- [134] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1505-1518, 2003.
- [135] G. V. Veres, M. S. Nixon, and J. N. Carter, "Modelling the time-variant covariates for gait recognition," *Proceedings of Audio- and Video-Based Biometric Person Authentication*, pp. 597-606, 2005.
- [136] S. Argyropoulos, S. D. Tzovaras, D. Ioannidis, M. G. Strintzis, N. V. Boulgouris, and Z. X. Chi, "A channel coding approach for human authentication from gait sequences," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 428-440, 2009.
- [137] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Factorial HMM and parallel HMM for gait recognition," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 39, no. 1, pp. 114-123, 2009.
- [138] X. Li, S. J. Maybank, S. Yan, D. Tao, and D. Xu, "Gait components and their application to gender recognition," *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 38, no. 2, pp. 145-155, 2008.
- [139] D. Ioannidis, D. Tzovaras, I. G. Damousis, S. Argyropoulos, and K. Moustakas, "Gait recognition using compact feature extraction transforms and depth information," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 623-630, 2007.

- [140] K. Bashir, T. Xiang, and S. Gong, "Feature selection for gait recognition without subject cooperation," *Proceedings of British Machine Vision Conference*, Leeds, U.K., September, 2008.
- [141] K. Bashir, T. Xiang, and S. Gong, "Feature selection on gait energy image for human identification," *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 985-988, 2008.
- [142] M.-L. Zhang, and Z.-H. Zhou, "ML-KNN: a lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038-2048, 2007.
- [143] Y. Zhang, and Z.-H. Zhou, "Multi-label dimensionality reduction via dependence maximization," *Proceedings of National Conference on Artificial Intelligence*, pp. 1503-1505, 2008.
- [144] R. E. Schapire, and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2-3, pp. 135-168, 2000.
- [145] F. D. Comite, R. Gilleron, and M. Tommasi, "Learning multi-label alternating decision tree from texts and data," *Lecture Notes in Computer Science*, vol. 2734, pp. 35-49, 2003.
- [146] A. Elisseeff, and J. Weston, "A kernel method for multi-labelled classification," *Advances in Neural Information Processing Systems*, vol. 14, pp. 681-687, 2002.
- [147] M. L. Zhang, and Z. H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338-1351, 2006.
- [148] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 57-68, 2007.
- [149] C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 572-581, 2004.

- [150] C. Liu, and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant method for face recognition, " *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467-476, 2002.
- [151] C. Liu, and H. Wechsler, "Independent component analysis of gabor feature for face recognition," *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 919-928, 2003.
- [152] X. Yang, Y. Zhou, T. Zhang, E. Zheng and J. Yang , "Gabor phase based gait recognition," *Electronic Letters*, vol. 44, no. 10, pp. 620-621, 2008.
- [153] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 755-761, 2009.
- [154] X. Yang, Y. Zhou, T. Zhang, G. Shu, and J. Yang, "Gait recognition based on dynamic region analysis," *Signal Processing*, vol. 88, no. 9, pp. 2350-2356, 2008.
- [155] J. Yang, J.-Y. Yang, D. Zhang, and J.-F. Lu, "Feature fusion: parallel strategy vs. serial strategy," *Pattern Recognition*, vol. 36, no. 6, pp. 1369-1381, 2003.
- [156] B. Xiao, X. Yang, Y. Xu, and H. Zha, "Learning distance metric for regression and semidefinite programming with application to human age estimation," *Proceedings of ACM International Conference on Multimedia*, pp. 451-460, 2009.
- [157] J.-H. Yoo, D. Huang, and M. K. Nixon, "Gender classification in human gait using support vector machine," *Proceedings of Advanced Concepts for Intelligent Vision Systems*, pp. 138-145, 2005.
- [158] S. H. Choa, J. M. Parkb, and O. Y. Kwona, "Gender differences in three dimensional gait analysis data from 98 healthy Korean adults," *Clinical Biomechanics*, vol. 19, no. 2, pp. 145-152, 2004.
- [159] J. Lu and E. Zhang, "Gait recognition for human identification based on ICA and fuzzy SVM through multiple views fusion," *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2401-2411, 2007.

- [160] J. Lu and Y.-P. Tan, "Uncorrelated discriminant simplex analysis for view-invariant gait signal computing," *Pattern Recognition Letters*, vol. 31, no. 5, pp. 382-393, 2010.
- [161] J. Lu and Y.-P. Tan, "A doubly weighted approach for appearance-based subspace learning methods," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 1, pp. 71-81, 2010.
- [162] J. Lu and Y.-P. Tan, "Doubly weighted nonnegative matrix factorization for imbalanced face recognition," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 877-880, 2009.
- [163] J. Lu and Y.-P. Tan, "Regularized locality preserving projections and its extensions for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 40, no. 3, pp. 958-963, 2010.
- [164] J. Lu and Y.-P. Tan, "Cost-sensitive subspace learning for face recognition," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2661-2666, 2010.
- [165] J. Lu and Y.-P. Tan, "Cost-sensitive subspace learning for human age estimation," *Proceedings of IEEE International Conference on Image Processing*, pp. 1593-1596, 2010.
- [166] J. Lu and Y.-P. Tan, "Uncorrelated discriminant nearest feature line analysis for face recognition," *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 185-188, 2010.
- [167] J. Lu and Y.-P. Tan, "Nearest feature space analysis for classification," *IEEE Signal Processing Letters*, vol. 18, 2011.
- [168] J. Lu and Y.-P. Tan, "Gait-based human age estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1718-1721, 2010.
- [169] J. Lu and Y.-P. Tan, "Gait-based human age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 761-770, 2010.