Dimensionality reduction methods and image-to-class distance for face recognition and human gait recognition

Huang, Yi

2010

Huang, Y. (2010). Dimensionality reduction methods and image-to-class distance for face recognition and human gait recognition. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/43807>

<https://doi.org/10.32657/10356/43807>

NANYANG TECHNOLOGICAL UNIVERSITY

Dimensionality Reduction Methods and Image-To-Class Distance for Face Recognition and Human Gait Recognition

A Thesis Submitted to
School of Computer Engineering
Nanyang Technological University

by

Huang Yi

in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy (Ph.D)

2010

Abstract

Biometrics has been a hot research topic in computer vision society in recent decades owing to its broad applications in commercial and government systems. Face and human gait are two of the most important biometrics which possess huge potential to recognize human unobtrusively. The whole recognition procedure can be divided into three steps: preprocessing, feature extraction/processing, and classification. This thesis focuses on the last two steps: feature extraction/processing and classification. For feature extraction/processing step, three different dimensionality reduction methods are proposed based on holistic features and local patch features respectively and a new distance metric is introduced in the classification step.

For the classification step, this thesis introduces an enhanced Image-to-Class distance to compute the distance from one probe image to the set of gallery images from the same person based on the local patch features. We formulate this task as an integer programming problem which incorporates the spatial constraint into each local patch feature by only allowing patches in a spatial neighborhood to be matched. Our proposed Image-to-Class distance is demonstrated to be more effective than the Image-to-Image distance and other existing distance measures in face and human gait recognition.

To deal with holistic features in the feature extraction/processing step, dimensionality reduction methods have been broadly used to enhance recognition performance. This thesis proposes the new Trace Ratio based Flexible Discriminant Analysis (TR-FSDA) which relaxes the hard constraint in Semi-supervised Discriminant Analysis (SDA) and adds a regularizer to better cope with data points sampled from non-linear manifold in face recognition. Then the objective function is further formulated in a trace ratio form instead of the conventional ratio trace form for better recognition rates. An iterative algorithm is proposed to compute the solution of the objective function, which is proved to converge to its optimal value based on the fast Newton-Raphson Method.

This thesis also employs local path features called the Patch Distribution Feature (PDF), in the feature extraction/processing step for both face and human gait recognition by using Gaussian Mixture Model (GMM) for describing the distribution of patch features. As PDFs are high-dimensional data, thus this thesis also proposes two dimensionality reduction methods in supervised setting and semi-supervised setting respectively. First, the Regularized Trace Ratio Discriminant Analysis (RTRDA) is proposed to map the PDF from high-dimensional spaces to low-dimensional spaces for better classifying samples from different classes in the supervised setting, which achieves the best Rank-1 recognition rate on the USF HumanID gait database compared with all the other methods proposed to date. Second, this thesis introduces the Patch Distribution Compatible Semi-Supervised Dimensionality Reduction (PDC-SSDR), which is tailored for the PDFs for both face and human gait recognition. After investigating the relationship between individual GMM components among different images, we add a new regularizer based on the sub-manifold smoothness assumption on individual GMM components in the objective function of Linear Laplacian Regularized Least Squares (LapRLS/L). Similar to TR-FSDA, we also relax the hard linear constraint to better cope with the nonlinear manifold data structure. The component-level label prediction matrix, the image-level label prediction matrix, the projection matrix and the bias term of our proposed objective function can be obtained simultaneously. PDC-SSDR significantly outperforms the state-of-the-art semi-supervised dimensionality reduction methods in both face and human gait recognition.

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor, **Prof. XU Dong**, for his mentoring and help throughout my research. His insights and ideas enlightened me in various aspects throughout the work. His instructions to the development of my academic thinking and writing as well as his diligent and serious attitudes towards research have been helpful for my current work and will be inspiring for my future work.

I would also like to thank my cosupervisor, **Prof. CHAM Tat Jen**, for his guidance in my research. I would like to pay my special thanks to **Prof. Michael BROWN**, who first brought me to the world of research. Additionally, I would express my appreciation to **Dr. Steve LIN** and **Prof. TAN Ping** for their supervision and collaboration when I was doing internship in Microsoft Research Asia (MSRA). I would also like to thank **Prof. Ivor TSANG** for his kind discussion and guidance.

I would like to express my thankfulness to the School of Computer Engineering of Nanyang Technological University for the financial support and Center for Multimedia and Networking Technology (CemNet) for the excellent working environment during my past years of study. Additionally, I would like to thank MSRA for providing me the Fellowship award and the internship opportunity.

During the past four years, my colleagues in CeMNet have provided me lots of help in my daily work. I would like to give my special thanks to my group peers, they're Dr. NIE Feiping, Mr. YANG Yi, Mr. DUAN Lixin, Mr. CHEN Lin, Dr. XU Yanwu, Mr. XU Xinxing, Mr. ZENG Zinan, Ms. HGUYEN Bich Hoang Anh, Mr. LI Wen, and Mr. LIU Yiming. With their helpful discussions and sincere encouragement, my PhD study has been fruitful and joyous. I would also like to express my great thankfulness to my lab mates, they are Dr. Desmond TSOI, Mr. Jimmy LEE, Mr. KOOT Poon Wei, Dr. ZHANG Yu, Dr. LI Jiaming, Ms. WANG Huan, Dr. HU Yiqun, et al, for their invaluable support.

Furthermore, I want to thank my boyfriend WANG Bin, for his love, support, understanding and encouragement, as well as his company throughout the good days and bad days of my life. Finally, I will pay my special gratitude to my dearest parents. Without their love and support, I could not finish the research in this thesis.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	ix
List of Tables	xi
List of Symbols	xiii
1 Introduction	1
1.1 Introduction on Biometrics	1
1.1.1 Face Recognition	3
1.1.2 Human Gait Recognition	5
1.2 Research Objectives and Contributions	7
1.2.1 Motivation	7
1.2.1.1 Dimensionality Reduction	7
1.2.1.2 Distance Metric	8
1.2.2 Main contribution	8
1.3 Thesis Organization	10
2 Literature Survey	13
2.1 Literature Survey on Face Recognition	14
2.1.1 Introduction to Face Recognition Databases	14
2.1.2 Holistic Feature Based Approaches	16
2.1.2.1 Dimensionality Reduction and Graph Embedding	16
2.1.2.2 Dimensionality Reduction Based on Tensor Representation	20
2.1.3 Local Feature Based Approaches	21
2.2 Literature Survey on Human Gait Recognition	23

2.2.1	Introduction to Human Gait Databases	24
2.2.2	Model Based Approaches	26
2.2.3	Appearance Based Approaches	27
2.3	Literature Survey on Distance Metrics	31
3	Face and Human Gait Recognition Using Image-to-Class Distance	33
3.1	Introduction	33
3.2	Image-to-Class Distance Based on Local Feature Representation	35
3.3	Experiments	41
3.3.1	Face Recognition	41
3.3.1.1	Database Description	43
3.3.1.2	Results and Discussion	43
3.3.1.3	Other Discussions	44
3.3.2	Human Gait Recognition	45
3.4	Conclusion	48
4	Trace Ratio Based Flexible Semi-supervised Discriminant Analysis	50
4.1	Introduction	50
4.2	Related Work	52
4.2.1	LDA and Trace Ratio Based LDA	52
4.2.2	Semi-supervised Discriminant Analysis (SDA)	54
4.2.3	Local and Global Consistency (LGC) and Gaussian Fields Harmonic Functions (GFHF)	54
4.2.4	Linear Laplacian Regularized Least Squares (LapRLS/L)	56
4.2.5	Flexible Manifold Embedding (FME)	57
4.3	Our Proposed Algorithm	57
4.3.1	Motivation	57
4.3.2	Our Objective Function and Solution	58
4.3.3	Discussion With the Existing Methods	63
4.4	Experiments	64
4.5	Conclusion	69

5	Patch Distribution Feature Based Methods for Face and Human Gait Recognition	70
5.1	Patch Distribution Features	70
5.1.1	Local Patch Feature Extraction	71
5.1.2	Universal Background Model	73
5.1.3	Maximum A Posteriori (MAP) Adaptation	76
5.1.4	Gallery Set Expansion	78
5.2	Regularized Trace Ratio Discriminant Analysis (RTRDA) with Patch Distribution Feature for Human Gait Recognition	78
5.2.1	Introduction	78
5.2.2	Proposed Algorithm	79
5.2.3	Simple Fusion	81
5.2.4	Experiments	81
5.2.5	Conclusion	84
5.3	Patch Distribution Compatible Semi-Supervised Dimensionality Reduction for Face and Human Gait Recognition	84
5.3.1	Introduction	84
5.3.2	Related Work	86
5.3.3	Patch Distribution Compatible Semi-supervised Dimensionality Reduction (PDC-SSDR)	87
5.3.3.1	Motivation	87
5.3.3.2	Proposed Objective Function	89
5.3.3.3	Solution	92
5.3.4	Experiments	95
5.3.4.1	Face Recognition	95
5.3.4.2	Human Gait Recognition	100
5.3.5	Conclusion	106
6	Conclusion	108
6.1	Summary of Contributions	108
6.2	Future Research Directions	111

References	114
List of Publications	127

List of Figures

1.1	Flowchart of recognition procedure and the relationship between algorithms proposed in this thesis.	11
2.1	Each row represents images within a gait cycle of a different person. For each row, the first seven images are the normalized and aligned binary silhouettes and the last image is the gray-level average image of the binary silhouettes.	29
3.1	Illustration of $N^l(i)$ (D^l is set to $\sqrt{2}$ in this example). For each feature p_i extracted at the i -th position (<i>i.e.</i> , pixel (u, v)) from probe image I_p , there are $9l_c$ candidate features (shown in yellow) from all gallery images $I_{g_1}, I_{g_2}, \dots, I_{g_{l_c}}$ of the c -th class.	35
3.2	Comparison of the Image-to-Image distance and the Image-to-Class distance (Distances are shown in rectangles). The images are all from the FERET database [1]. The Image-to-Image distances between the probe image and four gallery images from class Q_1 (correct class) are all very large, resulting in the probe image being wrongly classified (the nearest neighbor instance turns out to be from class Q_2). In contrast, the Image-to-Class distance between the probe image and class Q_1 (119.00) is smaller than the Image-to-Class distance between the probe image and class Q_2 (142.17), allowing for the probe image to be correctly classified.	37
3.3	Sample images cropped from the face database CMU PIE (top), FERET (middle) and FRGC (bottom) respectively. In the FERET database, from left to right the images are from the be, bd, bf, bg, ba, bc and bh categories.	42

4.1	Sample images cropped from three face databases Umist (first row), FERET (second row) and FRGC (third row).	65
4.2	Trace ratio values (<i>i.e.</i> , λ) of our objective function after different number of iterations on three databases. Our algorithm usually converges to the optimal values after less than eight iterations.	68
5.1	Real part of the Gabor wavelet kernels $\{\psi_{s,o} _{s=0}^{s=4} _{o=0}^{o=7}\}$	72
5.2	Sample images from the CMU PIE database [2] with four GMM components shown in red, green, blue and magenta. Circles of the same color are from GMM components with the same index. Please refer to the PDF file for better viewing.	87
5.3	Flowchart of the whole recognition process. Four steps of the process are indicated by red, blue, magenta and cyan rectangles respectively. These four steps are: 1) global GMM training (red rectangle); 2) MAP adaptation (blue rectangle); 3) PDC-SSDR (magenta rectangle); 4) label prediction (cyan rectangle).	91
5.4	Sample images cropped from face databases CMU PIE (first row), FERET (second row) and AR (third and fourth row) respectively. The images in the third row and fourth row are from two sessions of the same person in the AR database.	94
5.5	Performance variations (%) of Gray, Gray-SDA, Gray-LapRLS/L, PDF, PDF-SDA, PDF-LapRLS/L and PDC-SSDR on CMU PIE, FERET and AR face databases with respect to different values of K . The upper row shows the results for the unlabeled training data and the bottom row displays the results for the unseen test data. As Gray, Gray-SDA and Gray-LapRLS/L are not related to the PDF features, their recognition accuracies remain constant with respect to different values of K	98

List of Tables

2.1	Details of human gait databases in the literature summarized by [3]. . . .	25
2.2	Design of covariates tested in the probe sets in [3]. L–left camera, R–right camera, C–concrete surface, G–grass surface, A–first shoe type, B–second shoe type, BF–carrying a briefcase, NB–no briefcase, M–data collected in May, N–data collected in Nov.	26
3.1	Top-one recognition rates (%) of NN_Gray, NN, NFL, NBNN, Image-to-Image, and Image-to-Class on CMU PIE, FERET and FRGC databases.	44
3.2	Rank-1 recognition accuracy (%) comparison on the USF HumanID Gait database. For our Image-to-Class, the number in parentheses is the optimal D^l	49
3.3	Rank-5 recognition accuracy (%) comparison on the USF HumanID Gait database. For our Image-to-Class, the number in parentheses is the optimal D^l	49
4.1	Top-one recognition rates (mean $\pm$ standard deviation%) of LDA [4], TRDA [5], SDA [6], LapRLS/L [7], TCA [8], FME [9] and our TR-FSDA on the three databases. The results shown in boldface are significantly better than the others, judged by the t-test (with a significance level of 0.05).	67
5.1	Rank-1 and Rank-5 recognition rate (%) comparison on the USF HumanID Gait database.	82

5.2	Top-one recognition rates (%) of Gray, Gray-SDA, Gray-LapRLS/L, PDF, PDF-SDA, PDF-LapRLS/L, and PDC-SSDR on the CMU PIE, FERET and AR databases using <i>normalized</i> features. For each data set, the results shown in boldface are significantly better than the others, judged by the t-test (with a significance level of 0.05).	101
5.3	Top-one recognition rates (%) of DCT-PDF, DCT-PDF-SDA, DCT-PDF-LapRLS/L, and DCT-PDC-SSDR on the CMU PIE, FERET and AR databases. For each data set, the results shown in boldface are significantly better than the others, judged by the t-test (with a significance level of 0.05).	101
5.4	Rank-1 recognition accuracy (%) comparison on the USF HumanID Gait database. Two sets of results are reported. The first 12 algorithms are supervised algorithms from previous works, while the last 7 methods are algorithms discussed in this section with 'U' and 'T' denoting the results for unlabeled data and test data respectively.	103
5.5	Rank-5 recognition accuracy (%) comparison on the USF HumanID Gait database. Two sets of results are reported. The first 12 algorithms are supervised algorithms from previous works, while the last 7 methods are algorithms discussed in this section with 'U' and 'T' denoting the results for unlabeled data and test data respectively.	104

List of Symbols

SYMBOL MEANING

c	index of the class
C	number of classes
I_p	one probe image
I_{gr}	the r -th gallery image of the class
l_c	number of images in the c -th class
D^l	distance of l -neighborhood system
p_i	i -th patch in the probe image
q_j^r	j -th patch in the r -th gallery image
H_r	number of features in r -th gallery image in the class
G	number of features in the probe image
H	number of features in one gallery image
$N^l(i)$	index set of l -neighborhood system at position i
Q_c	the set of patches from gallery images of c -th class
$\hat{f}_{ij}$	the flow from feature p_i to q_j^c
$\tilde{X}$	vertex set

$\tilde{Y}$	vertex set
W	projection matrix
$\mathbf{m}$	feature dimension after dimensionality reduction
X_L	the matrix contains all labeled samples
x_i	feature vector of the i -th image
M_w	within-class scatter matrix
M_b	between-class scatter matrix
M_t	total scatter matrix
R_c	index set of labeled samples from the c -th class
$\hat{d}$	feature dimension
l	number of labeled training samples
u	number of unlabeled training samples
$\bar{x}_c$	mean of samples in the c -th class
$\bar{x}$	mean of all the labeled samples
$\mathbf{b}$	bias term
F	image-level label prediction matrix
F_i	the i -th row of matrix F
S_w	within-class similarity matrix
S_b	between-class similarity matrix
D_w	diagonal matrix
D_b	diagonal matrix
L_w	within-class Laplacian matrix
L_b	between-class Laplacian matrix

L_t	total Laplacian matrix
X	all labeled and unlabeled training samples
S	similarity matrix of both labeled and unlabeled training samples
D	diagonal matrix
L	Laplacian matrix of both labeled and unlabeled training samples
U	diagonal matrix
$\tilde{L}$	normalized Laplacian matrix both labeled and unlabeled training samples
$\tilde{L}_b$	augmented between-class Laplacian matrix
$\tilde{L}_t$	augmented total Laplacian matrix
$\tilde{L}_a$	intermediate matrix
α	predefined parameter
β	predefined parameter
$\tilde{I}$	augmented identity matrix
y_i	label of the i -th sample
Y	label matrix
$Y_{i.}$	the i -th row of the matrix Y
B	one type of Laplacian matrix
η_l	predefined parameter for labeled data
η_u	predefined parameter for unlabeled data
η	predefined parameter
η_∞	predefined large value parameter for labeled data
γ_A	predefined parameter
γ_I	predefined parameter

$\hat{I}$	columnly orthogonal matrix
t	index for iteration
λ_t	trace ratio value at t -th iteration
ξ	predefined threshold value
Z_t	intermediate matrix at t -th iteration
X_h	X coordinate at position h
Y_h	Y coordinate at position h
ω	dimension of local patch feature extracted from neighboring window
d	dimension of local patch feature with X-Y coordinates
$\psi_{s,o}$	Gabor kernel
s	scale of the Gabor kernel
o	orientation of the Gabor kernel
$\vec{x}$	vector of X-Y coordinates
N_1	height of the image
N_2	width of the image
N	number of patches from all labeled images
θ	parameter set
k	index of the component of the GMM
m_k	weight of the k -th component of the GMM
μ_k	mean vector of the k -th component of the GMM
Σ_k	covariance matrix of the k -th component of the GMM
K	number of components of the GMM
P	the set of all patches from all labeled training images

$\tilde{m}_k$	weight of the k -th component of the global GMM
$\tilde{\mu}_k$	mean vector of the k -th component of the global GMM
$\tilde{\Sigma}_k$	covariance matrix of the k -th component of the global GMM
ζ	weighting parameter
$\bar{\mu}_k$	expected mean vector of the k -th component of the adapted image-specific GMM
$\hat{m}_k$	weight of the k -th component of the adapted image-specific GMM
$\hat{\mu}_k$	mean vector of the k -th component of the adapted image-specific GMM
$\bar{l}_c$	number of labeled images of the c -th class after gallery expansion
$\bar{l}$	number of labeled images after gallery expansion
$\hat{\mu}_k^i$	k -th mean vector of the image-specific GMM of the i -th image
$\hat{m}_k^i$	k -th weight of the image-specific GMM of the i -th image
n_k	intermediate variable
v_k^i	k -th weighted mean of the image-specific GMM of i -th image
V_k	the set of the k -th weighted means
$\hat{\theta}_k$	the set of the k -th adapted weights and means of all training images
X_k^i	X coordinate of k -th component center in the i -th image
Y_k^i	Y coordinate at k -th component center in the i -th image
ϱ	index of images
ρ	index of GMM components
V_k	the set of k -th weighted means from all the training images
S^k	similarity matrix for k -th component of all training images
F^k	component level label prediction matrix for k -th component
ϵ	mean squared distance between neighboring pairs

S^k	similarity matrix for k -th components from all the training images
D^k	diagonal matrix computed from S^k
L^k	Laplacian matrix for k -th components from all the training images
$\hat{F}_i$	the set of component-level label prediction vectors of one image
z	number of nearest neighbors
τ	predefined parameter
ν	predefined parameter
Λ^k	diagonal matrix for k -th components from all the training images
A	intermediate matrix

Chapter 1

Introduction

1.1 Introduction on Biometrics

With the rapid development of computing theory and the surge of the computational capability of the modern computers, biometrics have attracted much attention in the past several decades and it will still be popular for the next several decades due to the huge demand by commercial and government enforcement.

The definition of biometrics is: the measures for uniquely identifying or recognizing humans based on one or more intrinsic physical or behavioral traits. In computer science, biometric is used to identify humans for access management, identity control, surveillance and human behavioral study. Compared with conventional identity measures for human (*e.g.*, identity card, passport, etc.), biometric has its unique advantage: it can not be replaced, eliminated or forged as the other forms, thus it is more reliable and secure.

The biometrics which can be successfully used to identify human should possess the following characteristics:

- Universality: each person should have the biometrics.
- Uniqueness: every person should have unique value measured by the biometrics.
- Permanence: it should remain the same over time.

- Collectability: it should be easy to obtain from the subject.
- Performance: it should be accurate and robust.

Biometrics can be divided into two main categories [10]:

- Physiological category: This type of biometrics are related to the shape of the human body, including fingerprint, face, DNA, Palm print, hand geometry, iris, retina, odor/scent, etc.
- Behavioral category: This type of biometrics is related to the behavior of a person, including typing rhythm, gait, voice, etc. Some researchers have combined the terms behavioral biometrics as behaviometrics for this category of biometrics.

Biometrics have been used in the applications such as homeland security, immigration check, personal device login, surveillance, and so on. Currently, the physiological biometrics, such as fingerprint, iris and retina, are very popular and they have been used in many applications. For example, at Walt Disney World, fingerprints are taken from the guests to ensure that the ticket is used by the authorized person. Although fingerprint, iris and retinal recognition systems are reliable, they require cooperative actions from the subject, which restrict their application in more challenging environments, such as airports. As a result, research on the more unobtrusive biometrics (*e.g.*, face, gait, etc.) received significant amount of attention in recent years. Face and human gait can be obtained more easily than fingerprint, iris or retina. Sometimes they can even be obtained without the subject's notice, which is very useful in surveillance applications. As face and human gait are usually captured and recognized based on images or videos, they have become very hot topics in computer vision research.

1.1.1 Face Recognition

Face recognition is one of the most successful applications of image analysis and understanding, owing to the huge demand from commercial market and government. The acquisition of a face image can be either noticeable to the subject or unobtrusive. Face can be recognized based on static photos or videos in either controlled or uncontrolled environments. The definition of an automatic face recognition system can be stated as follows: given still images or videos of a scene, recognize or verify one or more persons in the scene using a stored database of faces.

The first time a person's face image is captured by the recognition system is called "enrollment". During the enrollment, the face images from an individual are stored, and these images are called training images or gallery images. Subsequently, new face images (referred to as test images or probe images) are captured and compared with the images stored at the time of enrollment such that the person can be identified. For the constrained environment, the person should usually stand in front of the camera under specific illumination and pose. Recently, some research works have been proposed to handle face images in the unconstrained environment, where both the images stored in the database and the images for identification are from photos taken under arbitrary illuminations and poses.

There are two types of tasks that a face recognition system can conduct:

- Recognition: compare the captured test image with stored database images to answer the question "Who am I?". It is under the presumption that the individual for test should be a known person stored in the database.
- Verification: compare the captured test image against stored database images to answer the question "Am I who I say I am?". The individual is only verified as the known person if the comparison of the test image and the stored database images falls within a predefined threshold.

In this thesis, we only focus on the recognition task.

An automatic face recognition system needs to cope with the following major challenges, including illumination, expression, occlusion and pose variation. Conventionally, face areas are detected, cropped and scaled to a standard size before recognition. A lot of face detection methods have been proposed in the literature, i.e., [11, 12]. In some works, the faces from the images are even manually cropped [13, 14] to make sure that the faces are well aligned. After cropping the faces from the whole image, they are aligned according to the positions of the two eyes and normalized to a standard size with the two eyes located at predefined positions.

After the face images are aligned and normalized, features can be extracted from the images. Holistic features (*e.g.*, the concatenated gray-level pixels) can perform well when the images are well aligned. Usually, holistic features are with very high dimensions, and the total number of training samples from each subject is limited. Thus dimensionality reduction techniques are proposed to process the features by mapping them from high-dimensional spaces to low-dimensional spaces. Then, any classifier can be used to classify the samples in the low-dimensional spaces. In the literature, many dimensionality reduction methods have been proposed to find the best projection direction for classifying the training samples, and those methods have been widely applied for face recognition using holistic features. To cope with the illumination variations, different types of filters (*e.g.*, Gabor, etc.) are used to process the gray-level values. These filters often operate on small patches centered at each pixel, such that the neighborhood information can be used to represent the center pixel. The filtered images can be concatenated as holistic features, which are high-dimensional data. Dimensionality reduction methods can be applied to map the high-dimensional data into low-dimensional spaces for better computational and recognition performance. Apart from supervised and unsupervised methods, which only use the training data to learn the low-dimensional embeddings or projection directions, semi-supervised methods are also studied for face recognition.

Although face images are pre-aligned, misalignment can still be observed in most cases. So, local patch feature based approaches are proposed to alleviate the strong spatial constraint imposed by holistic features. The filtered image can also be used as the local patch features to reduce the undesired illumination variations. When an image is represented by a set of local patch features, it is desirable to develop new methods to effectively measure the distance between test images and training images. Different distance metrics are hence proposed for the local patch features and it is an interesting research direction. Apart from misalignment, local patch features combined with suitable distance measurements can also deal with expression, occlusion and pose variations to some extent.

1.1.2 Human Gait Recognition

As a behavioral biometric, the feasibility that gait can be used for human recognition was demonstrated in 1970s. Although physical pattern of bipedal locomotion is similar among different persons, human gait do vary from person to person. In 1976, psychological analysis by Johansson [15] demonstrated that human can distinguish different people by their gaits, which are expressed by the point light display placed at the joints of the body. Recently works showed that human can 1) distinguish human locomotion from other motion patterns; 2) recognize friends; and 3) recognize gender, direction of motion and carrying conditions [3]. The most recent psychological study further demonstrated that humans can indeed recognize people by their gaits [16]. Gait analysis has also been applied to clinical trials for patient management [17].

Compared with the widely used biometrics (*i.e.*, finger print, iris, retina, face, etc.), gait can be recognized from great distances. There are two types of information for characterizing and recognizing gait: (1) kinematics or dynamics of the body movement and (2) body appearance. Early works focusing on kinematics or dynamics try to fit

the movement into some models and the model parameters are used as the feature to represent the subject. The parameters can be dynamic parameters (*e.g.*, the stride length, joint angles of the legs and thighs, and speed) or static body parameters (*e.g.*, the size ratios of various body parts).

Recently, many research works have demonstrated that better performance can be achieved by directly using the body appearances for gait recognition. Usually, human gaits are captured by cameras from a distance for a period of time, which cover several cycles of the same gait pattern. To extract the appearance features from gait images, for each frame, the silhouette of the body is first extracted from the background, resulting in binary images with the region of pixels belonging to the moving person separated from the background. In the collection process of the gait database [3], simple techniques are used for background subtraction. The statistics of the RGB values in the background are computed beforehand and each pixel is defined as belonging to either human body or background by calculating its Mahalanobis distance to the estimated background value. The dynamic information of the body movement is also integrated into the temporal field of the binary silhouette sequence. Gait cycles can be easily detected from the binary silhouette images by computing the area of the lower-half part of the human body [3]. It is obvious that when two legs are apart, the area of the pixels in the lower-half body is the largest when the two legs are apart, while it is the smallest when two legs fold together at the swing stance. The background subtraction process may result in noisy binary silhouette images, thus, many research works proposed to alleviate this problem by averaging the binary silhouettes within one gait cycle to generate a gait energy image (GEI, also referred to as average silhouette image). Obviously, average silhouette images discard the temporal/dynamic information. However, they are less noisy and they can greatly reduce the computational complexity compared with the algorithms using the binary silhouette images. Therefore, every gait video clip can be extracted and computed

as a sequence of average silhouette images for recognition. Then new methods can be developed for gait recognition, which is similar to face recognition.

1.2 Research Objectives and Contributions

1.2.1 Motivation

In this thesis, we focus on appearance based face recognition and human gait recognition. Since they are both recognized based on the images, the same methods can be applied. As preprocessing steps are different from application to application, in this thesis we assume that all the training and test images are cropped and roughly aligned using existing automatic or manual techniques. Hence, we only focus on the feature extraction/processing and classification problems.

1.2.1.1 Dimensionality Reduction

The mostly widely used feature for face recognition and human gait recognition is the holistic feature concatenated by the pixel values of the whole image. In most cases, the holistic feature of every face or gait image is approximately of several thousand dimensions. On the other hand, only a few hundred training images are available from which the classifier or the statistical model can be learned. This leads to the so called curse of dimensionality and small sample-size problems, to which the dimensionality reduction methods are necessary. Dimensionality reduction methods can be regarded as the feature extraction step to extract more discriminant features before conducting classification.

Conventional dimensionality reduction methods employ the data distribution or manifold structure of the training samples to find the low-dimensional representation of the training data or the projection matrix for mapping the data. There are many types of

formulations to find the optimal embeddings or directions for mapping the data using different criteria. New dimensionality reduction algorithms can be developed by designing new criteria for better separating the data.

Based on the different information used in the recognition process, the techniques can be divided into three types: unsupervised, supervised and semi-supervised techniques. Unsupervised techniques only utilize the data distribution of the training samples without considering any label information. In supervised methods, label information is taken into consideration. Given the labeled and unlabeled data at the learning stage, semi-supervised methods can be applied to make use of both label information and the data structure information of labeled and unlabeled samples. For different settings of the application (supervised, unsupervised, semi-supervised), different types of dimensionality reduction methods can be developed.

1.2.1.2 Distance Metric

In most face and human gait recognition systems, very simple classifier (*i.e.*, Nearest Neighbor [18] (NN)) is employed for the classification. In NN, every test sample is classified as the class label of its nearest training sample. Thus, it is an interesting research topic to define the distance between the test sample and the training samples. The simplest distance is Euclidean distance (l_2 distance), which is used in many applications. However, it is computed based on element-wise subtraction, thus the recognition performances may be significantly degraded when misalignment exists. If local patch features are extracted from the images for classification, new distance metrics can be developed to effectively compute the distance between two images, represented by two sets of features. Misalignment can be taken into consideration in the development of the distance metric.

1.2.2 Main contribution

The following are the main contributions of this thesis:

- Based on local features, we propose a new classification method using the enhanced Image-to-Class distance. Each probe image (a face image or an average human silhouette image) is represented as a set of local features uniformly sampled over a grid with fixed spacing, and each gallery image is represented as a set of local features sampled at each pixel. Then we formulate an integer programming problem to compute the Image-to-Class distance from one probe image to all the gallery images belonging to the same person. We also enforce a spatial neighborhood constraint by only allowing neighboring features that are within a given spatial distance to be considered for feature matching. Image-to-Class distance is demonstrated to outperform the existing Image-to-Image distance and other state-of-the-art distance measures.
- For holistic features, we propose a new semi-supervised dimensionality reduction method called Trace Ratio based Flexible Semi-supervised Discriminant Analysis (TR-FSDA). In order to better cope with the data points sampled from nonlinear manifold, we first relax the hard constraint $F = X^T W$ by introducing a flexible regularizer $\|F - X^T W\|^2$ into the objective function of SDA. Then we further formulate our objective function in a trace ratio form. As there is no closed form solution to our new objective function, we propose an iterative algorithm to solve it and we prove that our iterative algorithm converges to the optimum based on the fast Newton-Raphson method.
- For local features, a new dimensionality reduction algorithm called Regularized Trace Ratio Discriminant Analysis (RTADA) is proposed in combination with the Gaussian Mixture Model (GMM) based Patch Distribution Feature for human gait recognition. We first extract Gabor features as the local patch features at every pixel of the average silhouette image and train a GMM to describe the distribution

of the patches in each image. A Universal Background Model (UBM) is first trained with local patch features from *all* gallery images, then every gallery or probe image is represented by the distribution parameters (referred to as Patch Distribution Features (PDF)) of the image-specific GMM adapted from the UBM. To cope with the high dimension of the PDF feature, the RTRDA is developed to find the most discriminant subspaces for gait recognition. We achieve the best Rank-1 recognition rate to date on the USF HumanID gait database, which is currently the largest human gait database.

- A new semi-supervised learning algorithm, referred to as Patch Distribution Compatible Semi-Supervised Dimensionality Reduction (PDC-SSDR), which is tailored for PDF, is proposed for both face and human gait recognition. Each image is first represented as a set of local patch features, which are further characterized as the corresponding Patch Distribution Feature. Assuming that the individual components of the GMM from all the training images reside on a sub-manifold, we assign component-level prediction labels for individual components and introduce a new regularizer based on the set of local sub-manifold smoothness assumptions in our objective function. We also constrain the component-level prediction labels to be consistent with the image-level prediction labels. Similar to TR-FSDA, we further integrate the regression residual $\|X^TW + \mathbf{1}b^T - F\|^2$ in our objective function to cope with data sampled from nonlinear manifolds. Finally, the optimal solutions of the component-level labels, the image-level labels, the projection matrix, and the bias term can be simultaneously obtained.

1.3 Thesis Organization

Figure 1.1 illustrates the flowchart of the procedure of face and gait recognition. A typical recognition procedure comprises three steps: preprocessing, feature extraction/process-

Figure 1.1: Flowchart of recognition procedure and the relationship between algorithms proposed in this thesis.

ing, and classification. In this thesis, four new algorithms are proposed in different steps of the recognition procedure. Chapter 3 proposes a new distance metric for classification step while Chapter 4 and Chapter 5 introduce three different dimensionality reduction algorithms for feature extraction/processing step based on holistic or local features in supervised or semi-supervised settings.

The remainder of this thesis is organized as follows:

In Chapter 2, we provide a literature survey about the previous works on face recognition and human gait recognition. For face recognition, we first introduce several benchmark databases used in this field, and then discuss the previous works in terms of holistic feature based approaches and local feature based approaches. For human gait recognition, we also introduce the databases at first and then review the literature in terms of model based approaches and appearance based approaches.

In Chapter 3, we introduce our work on Image-to-Class distance and its application to both face recognition and human gait recognition. We first extract local patch features

from each image and then compute the distance from one probe image to the set of gallery images from a certain class by solving an integer programming problem with spatial constraint.

In Chapter 4, the Trace Ratio based Flexible Semi-supervised Discriminant Analysis (TR-FSDA) is proposed for face recognition in the semi-supervised setting. We add a residue term into the objective function of Semi-supervised Discriminant Analysis (SDA) and further reformulate it in the trace ratio form. Then, an iterative algorithm is developed to acquire the optimal solution of the proposed objective function.

Chapter 5 presents two algorithms based on the Patch Distribution Feature (PDF), which is learned from the set of local patch features extracted from each image. Considering the PDF representation is of high dimensions, Regularized Trace Ratio Discriminant Analysis (RTRDA) is first proposed for dimensionality reduction for human gait recognition in the supervised setting. Then, we propose Patch Distribution Compatible Semi-Supervised Dimensionality Reduction (PDC-SSDR) for both face and human gait recognition. This algorithm is tailored for PDF by utilizing the characteristics of the GMM components in the development of the dimensionality reduction algorithm. Closed-form solutions can be obtained, and the predicted labels and projection matrix can be used for classification.

Chapter 2

Literature Survey

In this chapter, I will give a brief introduction to the two applications: face recognition and human gait recognition, as well as the recent research progress that has been made in these two areas. Since both face recognition and human gait recognition are performed on images or videos, some computer vision algorithms are applicable to both of them. A typical recognition system comprises three steps: preprocessing, feature extraction/processing, and classification. The preprocessing step consists of detection, cropping and normalization of the images. It should be noted that preprocessing step is crucial to face and human gait recognition, as most classification algorithms assume that the face images are roughly aligned. Without proper preprocessing step, even the best classification methods would have been useless since mostly they impose strong spatial constraints during classification. Actually image alignment is a challenging area which attracted research interests in recent years. In this thesis we skip the preprocessing step, and we will only focus on the feature extraction/processing and classification steps. I will review some important previous works related to feature extraction/processing, including the broadly used dimensionality reduction algorithms. In addition, we will also introduce a classification method based on new distance metric which can be used in classification step. From the survey of previous works, we can see that although significant progress has been made in these two areas, face recognition and human gait recognition are still

far from applicable in real world settings. Hence, there is plenty of room for research in feature extraction/processing and classification.

2.1 Literature Survey on Face Recognition

Face is one of the most important biometrics for machines to identify human and it has received huge amount of attention by computer vision society in recent decades.

As the result of the efforts which have been made to pursue the development of algorithms for face recognition, several face databases containing comprehensive face images have been established as the evaluation basis for the proposed algorithms. We will first introduce the benchmark face databases and then provide a survey on the recent proposed face recognition works.

Usually, face recognition systems can be divided into two categories based on the input: static images or videos. Video based methods can utilize more information than static image based methods, such as motion and multiple view angles, etc. Obviously, ignoring the temporal information in videos, the video based systems reduce to image based ones. As static images are more common as the input to recognition systems, in this thesis, I only focus on the static image based methods.

For feature extraction/processing step, both holistic features and local features are proposed for representing each face image in recent works. Based on different types of features, different methods are proposed. So in the following, I will give a literature survey mainly according to the type of features used in the recognition algorithm: holistic features or local features.

2.1.1 Introduction to Face Recognition Databases

I will introduce five benchmark face databases which are widely used in the literature: the FERET database, Umist database, CMU PIE database, FRGC database and AR database.

FERET Database The Face Recognition Technology (FERET) database [1] was collected between August 1993 and July 1996, and it contains 14126 images of 1199 individuals. Significant variations in pose, expression and illumination can be observed from the face images. Different poses are designed in the database with varying angles to the center of the camera. In many previous works as well as our papers, a subset of the FERET database, comprising 1400 images from 200 individuals (each with seven images from poses labeled *ba*, *bc*, *bd*, *be*, *bf*, *bg*, and *bh*) is used.

Umist Database The Umist [19] (Sheffield) Face database was released in 1998. It contains 575 gray-level images from 20 subjects taken under different poses from profile to frontal views. Different races, sexes and appearances can be observed in this database.

CMU PIE Database The CMU PIE (Pose, Illumination, and Expression) database [2] was collected between October 2000 and December 2000. It contains more than 40,000 facial images of 68 people. The images were acquired over 13 different poses, under 21 flashes with or without background lighting. So there are totally $21 \times 2 + 1 = 43$ different illumination conditions. Different facial expressions are observed in the images.

FRGC Database The Face Recognition Grand Challenge (FRGC) 1.0 database [20] was released on 5th May 2004, with 5658 images from 275 subjects (the number of images per subject varies from 6 to 48). Large variations in illumination, facial expression, presence of glasses, as well as occasional blur are present in the face images.

AR Database The AR face database [21] contains over 4,000 face images from 126 people (70 men and 56 women), which was created at the University of Alabama at Birmingham in 1998. All the images are of frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). Each person participated

in two sessions, which were separated by two weeks time. Around half of the images in each session are designed with different types of occlusions. In some of the images, large portions of the faces (sometimes even half of the faces) are occluded, making the database very challenging for recognition algorithms.

As pose, expression, illumination and occlusion are the major problems in face recognition systems, all the face databases are intentionally designed to contain images with large variations in these aspects, in order to test the robustness of the proposed algorithms.

2.1.2 Holistic Feature Based Approaches

Holistic feature based approaches usually take the concatenated gray-level values of all the pixels in the image as input. Since the holistic features are mostly with high dimensions, dimensionality reduction or graph embedding methods are applied to extract more discriminative features before recognition to enhance the computational efficiency as well as get better recognition results.

2.1.2.1 Dimensionality Reduction and Graph Embedding

Methods under Supervised or Unsupervised settings The most famous dimensionality reduction method is the eigenface method [22] which applies Principal Component Analysis (PCA) [18] to all the training data to find a projection matrix which retains minimal reconstruction error in terms of l_2 distance. PCA can be formulated as an eigenvalue decomposition problem. The optimal projection matrix computed by PCA is composed of the eigenvectors corresponding to the largest eigenvalues of the covariance matrix of the training samples. These eigenvectors correspond to the directions which encode most of the variances of the training data distribution, hence, after projecting all the data to the low-dimensional space, the most of the variances (energies) are preserved.

While PCA is an unsupervised algorithm which only utilizes the data structure without label information, the supervised method Fisher Discriminant Analysis (FDA) [18] takes the labels of the training samples into account to find an optimal projection direction. The goal of FDA is to maximize the scatter matrix formed by samples from different classes while minimize the scatter matrix formed by samples from the same class. Note that FDA can only solve for one direction, so it can be extended to its multi-dimensional form which is formulated as a trace ratio formulation. As there is no closed form solution for the trace ratio formulation, it is converted to a ratio trace form, referred to as Linear Discriminant Analysis (LDA) (also referred to as Fisherface [4]), which can be solved by generalized eigenvalue decomposition. Recently, several algorithms [23, 5, 24] have been proposed to solve the trace ratio form of LDA (also referred to as TRDA), and they are proved to outperform ratio trace based LDA in most cases by experiments in face recognition. Guo et al. [23] used a bisection method to iterative estimate the optimal value of the objective function, however, it can not get very accurate solution. Then the method proposed by Wang et al. [5] convertes the trace ratio problem into a trace difference problem, and then iteratively approaches the optimal value of the original trace ratio formulation. As the iterative process of [5] is very slow, in [24], a faster solution of the trace ratio problem is proposed based on the Newton-Raphson method.

The graph embedding methods such as Laplacian Eigenmap (LE) [25], Locally Linear Embedding (LLE) [26] and ISOMAP [27] can also be applied to face recognition. LE preserves the local neighborhood structure when producing the low-dimensional embeddings by enforcing the neighborhood relationship between data points in high-dimensional space is also preserved in low-dimensional space. LLE computes the low-dimensional embedding of each data such that any data can be well reconstructed by the linear combination of its neighbors. ISOMAP calculates the low-dimensional embedding of each data by preserving the pair-wise geodesic distances.

Neighborhood Preserving Embedding (NPE) [28] and Locality Preserving Projection (LPP) [29] are the linear approximations of LLE and LE, which also preserve the neighborhood information (neighbors in the original high-dimensional space should also be close in the low-dimensional space after projection). Laplacianfaces [30] uses LPP for face recognition, which tries to find a projection matrix to preserve the original topology of the face image data distribution to achieve good recognition accuracy.

In [31], a general framework, which is referred to as Marginal Fisher Analysis (MFA), is proposed and it is demonstrated that many dimensionality reduction and graph embedding algorithms (PCA, LDA, LE, LLE, ISOMAP, LPP, etc.) can be unified into one framework in which the desired statistical and geometric properties of the data are encoded as graph relations.

Other than using nearest neighbor or ϵ -neighborhoods [30] to construct the neighborhood graph, several new methods for constructing the graph are proposed recently. Discriminant Simplex Analysis (DSA) [32] constructs the graph relations by a new concept called K-nearest-neighbor simplex and then computes the projection matrix based on the graph. In [33], L1-Norm is used to provide better construction of the graph. Fu et al. proposed Correlation Embedding Analysis (CEA) [34], which incorporates both correlational mapping and discriminant analysis for better recognition rates. Recently some other works [35, 36] also introduce new discriminant analysis or subspace learning algorithms for face recognition.

Methods under Transductive or Semi-supervised Settings In PCA, LDA, LPP, LLE and ISOMAP, the projection matrix or the low-dimensional embedding is computed based on the training data, then every training or test data is mapped from high-dimensional space to low-dimensional space. However, there are some scenarios where large number of unlabeled data are also available at training stage, in which case a large

group of methods are proposed to use both the labeled and unlabeled data to learn a better projection matrix or low-dimensional representation of the data. In the transductive setting, all the test data are available as the unlabeled data at training stage, so no specific projection matrix is needed. Only low-dimensional representations of the labeled and unlabeled data are computed. Based on the low-dimensional representations, recognition can be performed by using any classifier. In the semi-supervised setting, apart from the unlabeled data available at training stage, there will be unseen test data for recognition after the training stage, thus a projection matrix is desired to be computed to map the unseen test data for recognition.

Transductive methods including CoTraining [37], Transductive SVM [38] and Manifold Regularization [39] try to predict the labels of the unlabeled data by using the labeled data and the distribution of both the labeled and unlabeled data. Label propagation methods LGC [40] and GFHF [41] propagate the labels from the labeled data to the unlabeled data by linear regression based on the manifold structure of all the data.

Semi-supervised learning methods [42, 6, 7, 8, 9] are proposed to learn a projection matrix from both the labeled and unlabeled data to map the data to low-dimensional space and recognition performance can be improved after the mapping. Manifold Regularization [7] adds a Laplacian geometric regularizer to the Least Squares (LS) and Support Vector Machine (SVM), resulting in Laplacian Regularized Least Squares (LapRLS) and Laplacian Regularized SVM (LapSVM). Semi-supervised Discriminant Analysis (SDA) [6] extends LDA [18] for semi-supervised setting by adding a geometrical regularizer of the data manifold. Flexible Manifold Embedding [9] extends both LGC/GFHF [40, 41] and LapRLS/L [7] by replacing the linear term and adding a residue term to the objective function of LapRLS/L to learn a better projection matrix. Note that the learned projection matrix is applicable to all the data, semi-supervised algorithms are more useful than the transductive ones because the input images for a face recognition system are usually not available at the training stage.

It is well known that in most cases transductive and semi-supervised algorithms outperform supervised and unsupervised ones because they utilize more information (*i.e.*, unlabeled data) for training and provide better projection directions or low-dimensional embeddings for recognition. Semi-supervised settings are also useful for face recognition since the labeled data require extensive human labor while the unlabeled data are abundant and easy to acquire.

2.1.2.2 Dimensionality Reduction Based on Tensor Representation

The dimensionality reduction methods reviewed in Section 2.1.2.1 all treat each image as a vector, which is a data point in high-dimensional space. Recently, some dimensionality reduction methods are proposed to represent each data point as a high-order tensor (an images can also be regarded as a 2-D tensor). In [43], the traditional PCA is extended to 2-D image and the experiments show that 2D-PCA achieves much better recognition rates than vector-based PCA on face recognition. Then Ye et al. [44] extended LDA to 2D-LDA which solves for two projection matrices iteratively and the experiments show that 2D-LDA achieves comparable performance as LDA while being much more computational efficient. 3-D tensor based methods [45, 13, 46] are also proposed for face recognition.

Instead of regarding each pixel as an element in vector based dimensionality reduction methods, tensor based methods take each row or each column of the image as an element, so the image structure is intrinsically preserved in the process. The number of parameters to be estimated is also reduced compared with the vector based methods, allowing the algorithms to learn parameters from small training sets, a situation which is usually encountered in face recognition. It is demonstrated that tensor based dimensionality reduction methods usually outperform vector based methods when there are very few training samples and the feature dimension is very high [45, 46].

2.1.3 Local Feature Based Approaches

Apart from the holistic feature based approaches briefed in Section 2.1.2, which employ the whole image as one data point in high-dimensional space, the local feature based approaches extract local features at selected positions of each face image and then the set of local features are used for recognition. Local features are more robust to misalignment and illumination, so a lot of local feature based methods have achieved good recognition results on face recognition.

Some early works utilize the geometry information of local features [47, 48] for graph matching, which has the potential to achieve invariance to affine transformations and local facial expressions. 1-D HMM [49] and pseudo-2D HMM [50] are also proposed for structural matching. Dynamic Link Architecture (DLA) is proposed by [51, 52] to define the face as a set of connected graph nodes. The most famous graph matching method for face recognition is the Elastic Bunch Graph Matching (EBGM) system [53, 54], which extracts 40-dimensional Gabor features from five different scales and eight different directions at each fiducial points (*e.g.* the pupils, the corners of the mouth, the tip of the nose, the top and bottom of the ears, etc.) to represent each facial component. EBGM expands DLA by representing each face image by a labeled graph and the graphs of test face images are compared with the training images by using a simple similarity function.

In EBGM, local Gabor features are extracted at some selected fiducial points on the face, while recently, some methods extract local patch features at every pixel of the face image and use the distribution of the local patch features to describe the image. The method proposed by Lucey and Chen [55] represents one face image as a set of local patch features. Then a global Gaussian Mixture Model (GMM) is trained from the local patch features extracted from *all* the training images as a Universal Background Model (UBM) to represent the population of faces. The UBM can be adapted to an image-specific GMM

based on the observations from one image [55] based on Maximum A Posteriori (MAP) probability. In [56], Lucey and Chen further extended their work to deal with mismatches caused by pose variation. The UBM proposed in [55] can ensure better fitness of the data to the GMM when there are not enough training data available. After the image-specific GMM is adapted for each image, the Patch Distribution Feature (PDF) is formed by the parameters of the image-specific GMM. Then recognition is performed on the PDF by using any classification algorithm. The GMM based methods can achieve good results on face recognition, however, the spatial information, which is crucial to this application, is abandoned by the local patch features. To overcome this problem, Liu et al. [57] extended [55] by adding the X-Y coordinates into the local patch features, enforcing the spatial alignment in face images as prior knowledge in model building process.

Local feature based methods are also popular in object recognition [58] and the Scale-invariant Feature Transform (SIFT) feature has achieved great success as a local patch feature for object recognition. Inspired by object recognition, SIFT feature has also been applied to face recognition in [59, 60]. Note that in object recognition, matching two sets of local features is a major issue. As face images are roughly aligned in the preprocessing step, slight misalignment is still visible in most images. Several methods have been proposed to cope with the misalignment problem by incorporating spatial constraint of each local patch feature. In [14], each image is represented as a set of local patch features extracted all over the image at a fixed spacing, then a non-parametric matching method is used to calculate the distances between two sets of local patch features from two images by using Spatially constrained Earth Mover's Distance (S-EMD). Similar to S-EMD, Zhou et al. [61] proposed to match each local patch feature of the probe face image to the features in the spatial neighborhood from the gallery images of a certain class. Each probe local feature is linearly reconstructed by the set of local features in the neighborhood from the gallery images of a certain class and the reconstruction error is used as the distance

for recognition. Obviously, in [14] and [61], matching within a spatial neighborhood can greatly reduce the misalignment effect. Wang et al. [62] proposed to compute the embedding of the data and the underlying spatial misalignment parameters on local patch features of the face image. In [63] and [64], 2-D and 3-D patch correspondences are learned to better cope with the viewpoint variances for face recognition in different poses.

How to design a good descriptor to represent the set of local patch features extracted from face images has become a popular research direction in recent years. Wright and Hua [65] proposed to use random projections to quantize the set of local patch features into different bins from which histograms are constructed and used as the descriptors for the images. Kumar et al. [66] learned a Volterra kernel which maps the local patch features to low-dimensional spaces and in which the intra-class distances are minimized while between-class distances are maximized. Cao et al. [67] proposed a learning based descriptor to characterize each image by quantizing the local patches by random projection tree and PCA tree recursively for each facial component. In [68], Cao et al. also adopted the component based method to build the face image retrieval system using local patch features. The local patch features are then quantized according to different persons to form descriptors.

2.2 Literature Survey on Human Gait Recognition

Gait is one of the biometrics that computer vision researchers try to investigate recently as a signature to identify human. It is a complicated dynamic behavior which is the combination of the movement of several body parts (*i.e.*, legs, arms, torso, and so on). It is affected by many factors such as shoes, surface of walking, clothes, carrying condition, body build and weight, emotional state, and so on. In the research works proposed by [69, 70, 71], experiments show that human gait can be identified by the movement of

the point-light placed at the joints of the body. Recently, more and more attentions have been attracted by human gait in computer vision society and a group of works have been done in this area [72, 73, 74, 75, 76, 77, 78] by modeling the body movement.

Human gaits are recorded by taking videos of people walking. From the videos, different types of gait features can be extracted and used for recognition. The existing methods for human gait recognition can be roughly divided into two groups: model based approaches and appearance based approaches. In model based approaches, the dynamic model parameters are extracted from the movement of the person in the video (*i.e.*, kinematics, stride length, joint angle of the legs, etc.), and the recognition is performed based on these parameters. On the other hand, in appearance based approaches, the silhouettes of human gait, which depict the shape and gesture of the body, are first extracted from the video, and then computer vision techniques are used to perform recognition based on these silhouettes.

In the following, we will first introduce the commonly used gait databases and then review the previous research works from both model based group and appearance based group.

2.2.1 Introduction to Human Gait Databases

There are mainly seven human gait databases proposed in the literature. The details of the databases are in Table 2.1, which are summarized mostly by [3].

Obviously, USF HumanID database is the largest human gait database to date, and it covers much more covariates compared with the other databases. Thus, in this thesis, we only focus on the USF HumanID database. USF HumanID gait database contains 122 people and each of them is required to walk on an elliptical path in front of different cameras under predefined circumstances, in order to test which covariates are more likely to affect the gait. In total, 1870 video clips from 122 subjects are collected covering five covariates: 1) viewpoints(left/right); 2) shoe types (A/B); 3) surface types (grass/concrete);

Table 2.1: Details of human gait databases in the literature summarized by [3].

Database	# Subjects	# Sequences	Scene	Data Covariates
UCSD 1998 [73]	6	42	Wall background	Time(minutes)
CMU MoBo 2001 [79]	25	600	Indoor, treadmill	Viewpoint, Walking speeds Carrying condition Surface incline
Georgia Tech 2001 [80]	15	168	Outdoor	Time(6 months)
	18	20	Magnetic tracker	Time(6 months)
Maryland 2001 [81]	25	100	Outdoor, 30m away	Surface incline
	55	222	Outside, Top mounted	Viewpoints (front, side) Time
MIT 2001 [82]	24	194	Indoor	Time(3 months)
Southampton 2001 [83]	28	112	Indoor, green background	Time(minutes)
CASIA Gait Dataset (A) 2001 [84]	20	240	Outdoor	3 Viewpoints Surface(concrete)
CASIA Gait Dataset (B) 2005 [84]	124	744	Indoor	11 Viewpoints Clothing Carrying condition
CASIA Gait Dataset (C) 2005 [84]	153	1530	Outdoor	Infrared images Walking conditions Carrying condition
USF HumanID 2002 [3]	122	1870	Outdoor	Viewpoint($\approx 30^\circ$) Surface(concrete, grass) Carrying condition Time(6 months)

Table 2.2: Design of covariates tested in the probe sets in [3]. L–left camera, R–right camera, C–concrete surface, G–grass surface, A–first shoe type, B–second shoe type, BF–carrying a briefcase, NB–no briefcase, M–data collected in May, N–data collected in Nov.

Exp.	Probe	# of	Difference
A	(G, A, L, NB, M/N)	122	View
B	(G, B, R, NB, M/N)	54	Shoe
C	(G, B, L, NB, M/N)	54	Shoe, View
D	(C, A, R, NB, M/N)	121	Surface
E	(C, B, R, NB, M/N)	60	Surface, Shoe
F	(C, A, L, NB, M/N)	121	Surface, View
G	(C, B, L, NB, M/N)	60	Surface, Shoe, View
H	(G, A, R, BF, M/N)	120	Briefcase
I	(G, B, R, BF, M/N)	60	Shoe, Briefcase
J	(G, A, L, BF, M/N)	120	View, Briefcase
K	(G, A/B, R, NB, N)	33	Time, Shoe, Clothing
L	(C, A/B, R, NB, N)	33	Surface, Time, Shoe, Clothing

4) carrying conditions (with/without a briefcase); 5) time of year (May/Nov, where the time covariate also implies different shoes and clothes). A gallery set is collected with all the 122 subjects walking on the grassland wearing shoe type A with no briefcase in May 2001. 12 probe sets are designed to test a single covariate or a set of covariates, containing different numbers of subjects varying from 33 people to 122 people. Table 2.2 (from [3]) shows the design of covariates for each probe set. Sarkar et al. [3] also extracted the binary gait silhouette images downloadable at <http://figment.csee.usf.edu/GaitBaseline/>.

2.2.2 Model Based Approaches

Kinematic models of the movement are learned from the training data and then parameters of the models are used for gait recognition.

Early work of [76] tries to model the physical structure of the human body by fitting a 2D skeleton to the body and estimating the joint angle of the thighs. In [85], a method is proposed to extract the coordinates of specific points on the body as the features of the gait and then gait is recognized by comparison of test sequences of coordinates with the

training sequences. In [80, 72], body skeleton parameters (*i.e.*, height, distance between head and pelvis, distance between pelvis and two feet) and the movement parameter (*i.e.*, the stride length) are extracted and used as the feature for gait recognition. In [86], the gait silhouette is divided into several parts, which are modeled by different sizes of ellipses. After the ellipses are fit into the body parts, the parameters of the ellipses are used for recognition. Tanawongsuwan and Bobick [87] estimated the time-normalized trajectories of joint angles in lower body by sensors placed at specific body positions as well as skeleton fitting. Then the trajectories are used as the feature in combination with Nearest Neighbor classifier [18]. Urtasun and Fua [88] proposed to fit 3-D temporal motion models to synchronized video sequences and then the models are used to track and recognize people. In [89], a hierarchical and structural model is proposed, which represents the legs by a set of thick lines to impose soft kinematic constraints. The time-varying parameters of the structural model are used to measure gait velocity, stance width, stride length, stance times, and other gait variables for recognition. Abdelkader [90] used stride and cadence, which are related to body height, weight, and gender, for recognition. In [91], a framework is proposed to track the walker and further recover joint-angle trajectories of lower limbs to form the dynamic cues. Then the static cues (*i.e.*, appearance of the upper body) and dynamic cues are fused on the decision level for recognition.

2.2.3 Appearance Based Approaches

Recently, appearance based approaches have become more and more popular in human gait recognition because they generally achieve better performance than the model based ones. In the appearance based approaches, the gait silhouette images are extracted from the video sequence. The silhouette images encode both the static information (body shape, height, etc.) and the dynamic information (stride, joint angle of legs, etc.)

of the gait. Static information comes from the shape and texture of the silhouettes while the dynamic information is incorporated in the temporal domain of the silhouette sequence. In the preprocessing step of the USF HumanID database [3], moving persons are detected and tracked from videos and then the binary silhouettes are extracted. On the USF HumanID database, a baseline algorithm [3] is proposed by computing the intersection of binary silhouettes between gallery and probe data. In [92], Liu and Sarkar proposed the Dynamics Normalized Gait Recognition (DNGR) which normalizes the gait dynamics using a generic walking model characterized by a population Hidden Markov Model (pHMM). The states of the pHMM represent different stances of one gait cycle and the silhouette images act as the observations for the corresponding stances of the cycle. Then the difference between two gait silhouette sequences of gallery and probe data are computed by the difference between two corresponding dynamics-normalized gait cycles. DNGR [92] achieves the best recognition results on USF HumanID database among existing methods, which was collected by the same authors. However, in [92], the pHMM is trained by using some manually cleaned data, which is not publicly available.

Although binary silhouettes encode the temporal information of the gait cycle, they are very noisy due to the error caused by detection and the noise from the environment. In [3], the gait cycles of each person are detected by using the periodical characteristic of the area of foreground pixels of the two legs. After the gait cycles are determined, the binary silhouettes of one gait cycle can be averaged to a gray-level image, which is called an average silhouette [93] or a Gait Energy Image (GEI) [94, 95]. A lot of recent works are done based on average silhouette images. There are three reasons for using average silhouette images instead of binary silhouette images: 1) the binary silhouette images extracted in [3] are noisy, which can be alleviated to some extent by average silhouette images; 2) average silhouettes will not be significantly changed if the starting stance of the cycle is different as long as the cycle is complete, so it is unnecessary to align

Figure 2.1: Each row represents images within a gait cycle of a different person. For each row, the first seven images are the normalized and aligned binary silhouettes and the last image is the gray-level average image of the binary silhouettes.

the binary silhouette images temporally; 3) the processing time can be greatly reduced by representing one cycle of binary silhouette images as an average silhouette image. Figure 2.1 shows two different people with each row representing the silhouette images within a gait cycle of one person from USF HumanID database. The first seven images are the normalized and aligned binary silhouette images within a cycle and the last image is the average gray-level image of the binary silhouette images.

Using average silhouette images for recognition, the temporal information are incorporated into each average silhouette image and human gait can be recognized by images, which is similar to face recognition. Hence, a lot of techniques for face recognition are also available for human gait recognition. Dimensionality reduction methods have been successfully used for human gait recognition. In [96], PCA is used on the silhouette images to reduce the dimensionality. Wang et al. [97] used procrustes shape analysis on the background subtracted silhouette images for gait recognition. In [94], Han and Bahnu proposed to apply PCA+LDA to the average silhouette images, and significant improvement from the baseline algorithm [3] can be achieved, demonstrating the effectiveness of dimensionality reduction methods for human gait recognition. Xu et al. [98] proposed a two stage algorithm which first removes the noise of the images by Concurrent Subspace

Analysis (CSA) [99] and then a tensor-based supervised algorithm Discriminant Analysis with Tensor Representation (DATER) [45] is applied for classification. In [100], a matrix-based extension of Marginal Fisher Analysis (referred to as MMFA) is proposed for human gait recognition and content-based image retrieval. The General Tensor Discriminant Analysis (GTDA), which is proposed in [101], first represents the silhouette images in the Gabor feature form, which is a high-order tensor representation, and then proposes a general tensor discriminant analysis to reduce the dimensionality of the data for better recognition rate. GTDA [101] achieves the best Rank-1 recognition results among all the algorithms that use the public USF HumanID database. Although the results of DNGR [92] are better than that GTDA [101], they are computed based on additional clean data which is not publicly accessible. Li et al. [102] proposed an extension of LLE [26], called Discriminant Locally Linear Embedding (DLLE) for high-order tensor data, which can be used for face and human gait recognition. Chen et al. [103] proposed another high-order tensor-based dimensionality reduction algorithm which preserves the pairwise geodesic distances of the nodes in the graph while enhances the discrimination ability of the data after projecting them to low-dimensional spaces.

Gender recognition based on human gait can be regarded as part of the task of gait recognition. Recently, Li et al. [104] proposed to cut the silhouette images into seven parts and perform recognition based on individual parts for gender recognition. Yu et al. [105] proposed to use the prior knowledge extracted from psychological experiments and conducted analysis on which human components are most discriminative for gender classification.

As the recognition performance on gait data alone may not be good enough to be used in real world applications, some research works focusing on fusion of gait and face data [106, 107] are proposed recently. This is undoubtedly advantageous than using gait or face data separately, however, acquisition of both face and gait data altogether is not

a trivial task. This is a new direction towards recognition applications in real world, hence, it definitely deserves more interests in the future.

2.3 Literature Survey on Distance Metrics

In many previous works of face and human gait recognition, after the global features are extracted for each image and dimensionality reduction is performed for mapping data to low-dimensional spaces, Nearest Neighbor (NN) Classifier [18] is used for classification. NN is widely used for its simplicity in implementation. The error rate of NN classifier is bounded at most twice as many errors as the optimal Bayes rule classifier [18], under the assumption that infinite training data are available. In the NN classifier, one probe image is classified to the class of its nearest gallery image, in which distance between two images is used for comparison. However, the NN classifier cannot generalize much beyond the labeled training data; hence the performance degrades acutely when the amount of training data is reduced. To overcome this problem, many new distance metrics have been proposed. The nearest feature line (NFL) method [108] is proposed to compute the distance between one probe image and each pair of gallery images belonging to the same class. By utilizing a linear model to interpolate and extrapolate each pair of gallery images from the same class, NFL provides more virtual samples of each class. NFL is further extended to nearest feature plane (NFP) and nearest feature space (NFS) [109], in which the distance from one probe image to the plane or the space constructed by samples from a certain class is computed for classification. Other than using individual images to compute distances for classification, a group of images from the same person are used as a whole set to compute the distances. Wang [110] used manifold-manifold distances to compute distance between image sets of different people for face recognition.

On the other hand, if there exists misalignment in the image, NN may not perform well using holistic features. Hence, other types of distance metrics have been proposed

to compute the distances between sets of local features. The Image Euclidean Distance (IMED) is proposed by Wang et al. [111], which computes the distance between two images based on the spatial relationship of neighboring pixels. It is demonstrated that IMED can better cope with misalignment in the image for face recognition [111]. In [101], IMED is also proved by experiments to outperform the baseline algorithm of NN. The Earth Mover's Distance (EMD) is widely used in object recognition to measure the distance between two sets of different numbers of local features [58]. The S-EMD method [14] reviewed in Section 2.1.3 is also a new distance metric which extends the EMD by enforcing the spatial constraint on the local features extracted from each image. Bioman et al. [112] proposed the distance from one test image to the set of training images from a certain class for object recognition and it is demonstrated to be useful for local feature based methods.

Chapter 3

Face and Human Gait Recognition Using Image-to-Class Distance

In this chapter, we introduce the Image-to-Class distance based on local feature representation for both face and human gait recognition.

3.1 Introduction

The nearest neighbor (NN) classifier is frequently used in face recognition and human gait recognition. In the NN classifier, one probe image is classified to the class of its nearest gallery image, in which distances between pairs of images are used for comparison. However, the NN classifier cannot generalize much beyond the labeled training data. When there is not enough training data to describe the whole data distribution, the performance of NN will significantly deteriorate. To improve the performance of the NN classifier, the nearest feature line (NFL) method [108] composes virtual samples by expanding each pair of gallery images of the same class from two points in the feature space into a line. Then the distance between one probe image and each line is computed for recognition. NFL is further extended to nearest feature plane (NFP) and nearest feature space (NFS) in [109], in which more virtual samples are produced by composing planes or spaces by gallery samples from the same class.

In the above classification methods NN, NFL, NFP and NFS, each image is generally represented as a holistic appearance feature (*e.g.*, a vector of gray-level pixels). But face images or human silhouette images may not be correctly aligned in the preprocessing step. In face recognition, recent studies have shown that the positions of the landmarks (*e.g.*, two eyes) from human annotation or an automatic face alignment algorithm may not be precise [113], which result in spatially misaligned face images (*i.e.*, the pixels for the same facial landmarks of the cropped images may not be strictly matched). In human gait recognition, it is still an open problem to determine the landmarks for silhouette images. Thus spatial misalignment also exists in human gait recognition because simple scaling and centering operations cannot precisely align silhouette images without landmarks. As shown in a recent study, spatial misalignment can significantly degrade the performance of holistic appearance feature based approaches [114].

To deal with spatial misalignment in face recognition and human gait recognition, we propose a new non-parametric classification method based on local feature representation. We represent each probe image (a face image or an average human silhouette image) as a set of local features uniformly sampled over a grid with fixed spacing, and characterize each gallery image as a set of local features sampled at each pixel. We formulate an integer programming problem to compute an *Image-to-Class distance from one probe image to all the gallery images belonging to a certain class*, in which any feature in the probe image can be matched to only one feature from one of the gallery images. Our Image-to-Class distance is motivated from two aspects: 1) the prior work NFL and its extensions have experimentally demonstrated better face recognition performance by using *the distances between a probe image and two or three gallery images from a certain class*, when compared with the nearest neighbor classifier based on Image-to-Image distance; 2) researchers have also demonstrated that it is useful to infer new image configurations by composing pieces from a set of other images for several different applications [115, 112], such as action classification, speaker recognition and retrieval.

Figure 3.1: Illustration of $N^l(i)$ (D^l is set to $\sqrt{2}$ in this example). For each feature p_i extracted at the i -th position (*i.e.*, pixel (u, v)) from probe image I_p , there are $9l_c$ candidate features (shown in yellow) from all gallery images $I_{g_1}, I_{g_2}, \dots, I_{g_{l_c}}$ of the c -th class.

Considering that face images or average human silhouette images are roughly aligned in the preprocessing step, we enforce a spatial neighborhood constraint for computational efficiency and effectiveness, namely, only the neighboring features that are within a given distance in the spatial domain are considered for feature matching. To efficiently cope with the integer programming problem, we treat it as a minimum-weight bipartite graph matching problem and apply the Kuhn-Munkres algorithm [116] to solve it. The comprehensive experiments on three benchmark face databases PIE, FERET and FRGC as well as the USF HumanID gait database clearly demonstrate the effectiveness of our newly proposed Image-to-Class distance. This chapter is structured as follows. In Section 3.2, we describe our Image-to-Class distance. The experimental results on face recognition and human gait recognition are presented in Section 3.3. We conclude our work in Section 3.4.

3.2 Image-to-Class Distance Based on Local Feature Representation

Similarly as in [14], we represent one probe image I_p as a set of local features $I_p = \{p_1, p_2, \dots, p_G\}$, where $p_i \in \mathbb{R}^d$ represents the feature (*e.g.*, a Gabor feature) extracted at the i -th position, and G is the total number of features. We denote all the gallery images

in the c -th class as $\{I_{g_1}, I_{g_2}, \dots, I_{g_{l_c}}\}$, where l_c is the total number of images in the class. Each gallery image is similarly represented as a set of local features $I_{g_r} = \{q_1^{g_r}, \dots, q_{H_r}^{g_r}\}$, $r = 1, 2, \dots, l_c$, where $q_j^{g_r}$ ($j = 1, 2, \dots, H_r$) is the feature extracted at the j -th position and H_r is total number of features extracted in the r -th gallery image of the class. For the probe image, we uniformly extract features over a grid with fixed spacing (say three pixels in this work). For gallery images, we densely extract features at every pixel. Thus we have $H_1 = H_2 = \dots = H_{l_c} = H$. Suppose the size of the gallery image is 36×36 , then G and H are 144 and 1296 respectively. While it is possible to densely extract features for probe image I_p at every pixel, the algorithmic efficiency will be significantly degraded because of the increase of the total number of features for I_p . Our experiments demonstrated that we can achieve comparable recognition performance when the spacing is set to three pixels.

We denote $Q_c = \{q_1^c, q_2^c, \dots, q_{H \times l_c}^c\} = \{q_1^{g_1}, \dots, q_H^{g_1}, q_1^{g_2}, \dots, q_H^{g_2}, \dots, q_1^{g_{l_c}}, \dots, q_H^{g_{l_c}}\}$ as all the features extracted from all the gallery images in the c -th class. We also define $g(p_i, q_j^c) = \|p_i - q_j^c\|$, $i = 1, 2, \dots, G, j = 1, 2, \dots, H \times l_c$, as the Euclidean distance between feature p_i and feature q_j^c . For each feature p_i in image I_p , we denote $N^l(i)$ as the index set of l neighboring features in the spatial domain from all the gallery images $I_{g_r}|_{r=1}^{l_c}$ in the c -th class, bounded by a distance D^l . Figure 3.1 shows an example for $N^l(i)$, in which D^l is set to $\sqrt{2}$. For each feature p_i extracted at the i -th position from the probe image I_p , there are $9l_c$ candidate features¹ in $N^l(i)$ for consideration. The Image-to-Class distance from the probe image I_p to all the gallery images in the c -th class is defined as:

$$d(I_p \rightarrow Q_c) = \sum_i \sum_{j \in N^l(i)} \hat{f}_{ij} g(p_i, q_j^c) / \sum_i \sum_{j \in N^l(i)} \hat{f}_{ij}, \quad (\text{Eq. 3.1})$$

where $\hat{f}_{ij}$ represents the flow from the feature p_i to q_j^c . In this work, the flow $\hat{f}_{ij}$ is restricted to binary values, specifically, if feature p_i and feature q_j^c are matched, then

¹The total number of candidate features is less than $9l_c$ for border pixels.

Figure 3.2: Comparison of the Image-to-Image distance and the Image-to-Class distance (Distances are shown in rectangles). The images are all from the FERET database [1]. The Image-to-Image distances between the probe image and four gallery images from class Q_1 (correct class) are all very large, resulting in the probe image being wrongly classified (the nearest neighbor instance turns out to be from class Q_2). In contrast, the Image-to-Class distance between the probe image and class Q_1 (119.00) is smaller than the Image-to-Class distance between the probe image and class Q_2 (142.17), allowing for the probe image to be correctly classified.

$\hat{f}_{ij} = 1$; otherwise $\hat{f}_{ij} = 0$. Therefore, Eq. 3.1 aims to calculate the average distance from the matched features. $\hat{f}_{ij}$ is optimally computed through solving the following integer programming problem:

$$\begin{aligned}
 \hat{f}_{ij} = \underset{f_{ij}}{\operatorname{argmin}} \quad & \sum_i \sum_{j \in N^l(i)} f_{ij} g(p_i, q_j^c), \\
 \text{s.t.} \quad & \\
 1. \quad & f_{ij} \in \{0, 1\}, \forall i, j, \\
 2. \quad & \sum_{j: j \in N^l(i)} f_{ij} = 1, \forall i, \\
 3. \quad & \sum_{i: i \in N^l(j)} f_{ij} \leq 1, \forall j.
 \end{aligned} \tag{Eq. 3.2}$$

Discussion: 1) From Eq. 3.2, we observe that each feature p_i from the probe image I_p will be matched to a corresponding feature extracted from neighboring positions in

one of the gallery images belonging to a certain class. In multi-view face recognition, some parts of an image A captured by a frontal-view camera will not be matched to corresponding parts from an image B captured by a left profile-view camera, due to self-occlusion. The same situation applies when comparing image A to image C captured by a right profile-view camera. On the other hand, it is much more likely that each feature of image A can be matched to either a feature in image B or in image C. More specifically, while the Image-to-Image distance from image A to B and from image A to C are both large, it is possible that the Image-to-Class distance from image A to a class comprising the two images B and C is small.

In Figure 3.2, we compare a probe image with all the gallery images from two different classes using Image-to-Image distance and Image-to-Class distance. The true label of the probe image is class Q_1 , but the Image-to-Image distances between the probe image and four gallery images from class Q_1 are all very large, making the probe image wrongly classified (the nearest neighbor is from class Q_2). In contrast, the Image-to-Class distance between the probe image and class Q_1 (119.00) is smaller than the Image-to-Class distance between the probe image and class Q_2 (142.17), allowing for the probe image to be correctly classified.

2) As shown in Eq. 3.2, each probe feature p_i is required to be matched to a gallery feature. Conversely, while each gallery feature q_j^c is restricted to match at most one probe feature, the distance measure allows for unmatched gallery features. This relates to the restriction of flow $\hat{f}_{ij}$ to binary values.

3) Since the face images or averaged human silhouette images used are to be roughly aligned in the preprocessing step, we enforce a spatial neighborhood constraint in Eq. 3.2 for robustness and computational efficiency. If the size of the image is 36×36 pixels, the total number of variables is limited to $144 \times 9l_c$ if D^l is restricted to $\sqrt{2}$. Without a spatial neighborhood constraint, the total number of variables expands to $144 \times 1296l_c$, which

is computationally prohibitive. In addition, with the spatial neighborhood constraint, it prevents improbable matching of features, say from the upper-left image area to the bottom-right image area.

Solution: The above integer programming problem in Eq. 3.2 can be treated as an assignment problem, to which we pursue a minimum-weight matching solution in a bipartite graph. In our bipartite graph, we denote two vertex sets as $\tilde{X}$ and $\tilde{Y}$, where each vertex in $\tilde{X}$ represents one feature in the probe image I_p and each vertex in $\tilde{Y}$ represents an element in Q_c (*i.e.*, one feature from any gallery image belonging to the c -th class). If the feature p_i in vertex set $\tilde{X}$ is matched to the feature q_j^c in vertex set $\tilde{Y}$, we connect p_i with q_j^c with an edge of weight $g(p_i, q_j^c)$ (*i.e.*, the cost of feature matching). If p_i and q_j^c are connected, we have $\hat{f}_{ij} = 1$; otherwise, $\hat{f}_{ij} = 0$. In addition, there should be no edges connecting two vertices from the same set, namely, the graph $\tilde{G} = (\tilde{X}, \tilde{Y})$ is a bipartite graph. In this work, we use the classical Kuhn-Munkres algorithm to solve the minimum weight bipartite graph matching problem. If the total number of elements in $\tilde{X}$ and $\tilde{Y}$ are both H , the complexity of the Kuhn-Munkres algorithm is H^3 .

Algorithm 3.1 Image-to-Class Distance Algorithm.

Given one probe image I_p and all the gallery images.

Represent the probe image I_p as a set of local features $I_p = \{p_1, p_2, \dots, p_G\}$, where $p_i \in \mathbb{R}^d$ represents the feature (*e.g.*, a Gabor feature) extracted at the i -th position.

Denote the gallery images in the c -th class as $\{I_{g_1}, I_{g_2}, \dots, I_{g_{l_c}}\}$. Calculate the Image-to-Class distance $d(I_p \rightarrow Q_c)$ from the probe image I_p to all the gallery images in the c -th class.

- Represent each gallery image in the c -th class as a set of local features $I_{g_r} = \{q_1^{g_r}, q_2^{g_r}, \dots, q_H^{g_r}\}$, $r = 1, \dots, l_c$.
- Solve the optimal binary flow in Eq. 3.2 by using the Kuhn-Munkres algorithm.
- Calculate the Image-to-Class distance $d(I_p \rightarrow Q_c)$ by using Eq. 3.1.

After calculating the Image-to-Class distances $d(I_p \rightarrow Q_c)$ to all the classes, perform classification by using Eq. 3.3 and output the optimal class c^* .

Classification: For each probe image I_p , we compute the Image-to-Class distances $d(I_p \rightarrow Q_c)$ to all the classes, and choose the optimal class c^* that has the minimum Image-to-Class distance, *i.e.*,

$$c^* = \arg \min_c d(I_p \rightarrow Q_c). \quad (\text{Eq. 3.3})$$

The detailed procedure of our method is given in Algorithm 3.1.

Relation to prior work: The prior work NFL and extensions NFP and NFS considered the distances between one probe image to two or more gallery images from a certain class. However, these works represent each image as a holistic appearance feature, and therefore they are sensitive to spatial misalignment. In contrast, our work represents an image as a set of local features and allow features extracted at different grid positions to be matched, which is more robust to spatial misalignment.

The prior work [14], referred to as Spatially constrained Earth Mover’s Distance (S-EMD), also coped with spatial misalignment in face recognition by using local feature representation. However, the classification in [14] is based on the Image-to-Image distance between one probe image and a single gallery image. Moreover, in this work we employ the Kuhn-Munkres algorithm based on bipartite graph matching to solve the integer programming problem in Eq. 3.2, which is of substantially lower computational complexity than the simplex solution used in [14].

Recently, Boiman *et al.* proposed a Naive-Bayes NN image classifier (NBNN) for general image classification [112]. While NBNN also aims to compute the distance between one query image and several database images belonging to a certain class, NBNN used the classical bag-of-words model without considering any spatial information. For general image classification, it is conventional to represent one image as a set of orderless local SIFT descriptors [117], possibly to bypass the open problem of aligning general images. In contrast, while some spatial misalignment exists, face images or human silhouette images are *roughly* aligned in the preprocessing step. Thus, it is beneficial to

utilize approximate spatial information in these situations. To this end, we enforce an effective spatial constraint and formulate an integer programming problem to compute the Image-to-Class distance. In addition, if we directly apply NBNN for face recognition or human gait recognition, two different features of the probe image may be matched to the same feature of one gallery image. As an intuitive example, it is evident that two eyes of a face in the probe image should not be matched to the same eye of a gallery image. In contrast, for any feature in the gallery images, at most one feature from the probe image I_p can be matched to it.

3.3 Experiments

We compare our proposed Image-to-Class distance (referred to as *Image-to-Class*) to several state-of-the-art algorithms for face recognition and human gait recognition. We choose three benchmark face databases: CMU PIE [2], FERET [1], FRGC [20], as well as the USF HumanID gait database [3].

3.3.1 Face Recognition

For face recognition, we align all the gray-level images by fixing the locations of the two eyes and normalizing the cropped images to 36×36 pixels, followed by histogram equalization. For the CMU PIE and FERET databases, the locations of the two eyes are manually labeled. For the FRGC database, we utilize the Active Shape Model [11] as the face alignment algorithm to automatically locate the key points on the face. Some example images after preprocessing are shown in Figure 3.3.

Each probe image I_p is represented as a set of local features $I_p = \{p_1, p_2, \dots, p_{144}\}$, where p_i represents a 9-dimensional gray-level feature vector extracted from the surrounding 3×3 patch centered at the i -th position. Each gallery image I_{g_r} is represented as a set of local features $I_{g_r} = \{q_1^{g_r}, q_2^{g_r}, \dots, q_{1296}^{g_r}\}$, $r = 1, \dots, l_c$, where $q_j^{g_r}$ is similarly

Figure 3.3: Sample images cropped from the face database CMU PIE (top), FERET (middle) and FRGC (bottom) respectively. In the FERET database, from left to right the images are from the be, bd, bf, bg, ba, bc and bh categories.

defined. To reduce the effect of illumination variation, we normalize p_i and q_j^{gr} to have zero mean and unit variance. We refer to p_i and q_j^{gr} post-normalization as normalized local features.

We compare our Image-to-Class distance measure to the NN [118], NFL [108] and NBNN [112] methods using the normalized local features. For the NN and NFL methods, each probe or gallery image is represented as a 11,664-dimensional feature by putting the 9-dimensional normalized features from all pixels together ($11664 = 1296 \times 9$). We also report results of directly applying the NN classifier based on the original 1296-dimensional gray-level feature (referred to as *NN_Gray*). In addition, we compare this work to the prior work *Image-to-Image* matching method in [14], in which the distance between one probe image I_p and a single gallery image I_{gr} is calculated with Eq. 3.2. In this case, $N^l(i)$ in Eq. 3.2 is the index set of l neighboring features in the spatial domain from image I_{gr} , and each probe image is labeled with the class of its nearest gallery image. For the Image-to-Class and Image-to-Image methods, we fixed D^l to $2\sqrt{2}$ for all the databases. Note that there is no other parameter in our Image-to-Class algorithm.

3.3.1.1 Database Description

The CMU PIE (Pose, Illumination, and Expression) database [2] contains more than 40,000 facial images of 68 people. The images are acquired over different poses, under variable illumination conditions, and with different facial expressions. Five near-frontal poses ($C27$, $C05$, $C29$, $C09$ and $C07$) and four illumination conditions (08, 11, 10 and 13) are chosen in this experiment. Due to the incompleteness of data, only 63 subjects are used in the experiments (each with twenty images). The images are partitioned into gallery and probe sets. In this work, six images per person are randomly selected for training and the remaining fourteen images are used for testing.

We also test our algorithm on a subset of the FERET [1] database, comprising 1400 images from 200 individuals (each with seven images labeled as ba , bc , bd , be , bf , bg , and bh under different views with angles $\{0^\circ, +40^\circ, +25^\circ, +15^\circ, -15^\circ, -25^\circ, -40^\circ\}$ to the center). In this work, four images from bd , be , bf and bg are used for training and the remaining three images from ba , bc and bh are used for testing.

We also use the FRGC1.0 database [20], containing 5658 images from 275 subjects (the total number of images per subject varies from 6 to 48). We observe large variability in illumination, facial expressions, presence of glasses, as well as occasional blur. Four images per person are randomly selected for training and the remaining images are used for testing. Please see Section 2.1.1 for more details of these face databases.

3.3.1.2 Results and Discussion

The experiments are shown in Table 3.1. We make the following observations:

- (i) We observe that NN outperforms *NN_Gray*, which demonstrates that the normalized local feature is an effective feature for face recognition.
- (ii) NFL [108] outperforms NN on the CMU PIE and FRGC databases, which demonstrates that piecewise linear extension of the sample set in NFL improves face

Table 3.1: Top-one recognition rates (%) of NN_Gray, NN, NFL, NBNN, Image-to-Image, and Image-to-Class on CMU PIE, FERET and FRGC databases.

Algorithm	CMU PIE	FERET	FRGC
NN_Gray [118]	48.8	50.8	54.2
NN [118]	78.0	53.5	64.9
NFL [108]	83.3	52.8	68.1
NBNN [112]	57.9	47.0	43.4
Image-to-Image [14]	84.1	88.8	69.1
Image-to-Class (This work)	90.4	90.0	71.0

recognition performance. However we also observe that doing so degrades performance in some cases. For example, NFL is slightly worse than NN on the FERET database.

- (iii) While NBNN [112] considers the distances between one probe image to multiple gallery images of a certain class, the performance of NBNN is worse than the basic NN classifier [118] on all the three face databases, possibly because the crucial spatial information for face images is discarded in NBNN.
- (iv) By coping with spatial misalignment, the Image-to-Image method [14] outperforms the holistic appearance based approaches of NN and NFL in all the three databases. By further utilizing the distances from one probe image to all the gallery images of a certain class, the Image-to-Class method proposed in this chapter achieves the best results in all three face databases.

3.3.1.3 Other Discussions

An experiment is carried out on a slightly different implementation of our Image-to-Class distance measure (referred to as *Image-to-Class-DenseProbe*), in which features are densely extracted at every pixel for the probe images, *i.e.*, each probe image I_p is represented as set of 1296 local features $I_p = \{p_1, p_2, \dots, p_{1296}\}$. The experiment is

carried out on the CMU PIE database. We observe that the recognition accuracy for the *Image-to-Class-DenseProbe* measure on the CMU PIE database is 92.0%, which is better than the result of 90.4% (see Table 3.1) for the *Image-to-Class* distance measure where the features from the probe images are sampled on a grid with 3-pixel spacing. However, we also observed that the average processing time for computing the distance from one probe image to six gallery images belonging to a certain class significantly increases from 0.024 seconds to 1.148 seconds on an IBM workstation (2.13 GHz CPU with 16 Gbyte RAM). In order to have a realistic tradeoff between algorithmic efficiency and effectiveness, we chose to continue using the *Image-to-Class* measure rather than the *Image-to-Class-DenseProbe* measure in subsequent experiments.

We also compare our Image-to-Class method to a variant of NBNN (referred to as NBNN-Var), in which each feature p_i in image I_p can only be matched to neighboring features in the spatial domain, bounded by a distance D^l . For both methods we set $D^l = 2\sqrt{2}$. Unlike our Image-to-Class method, NBNN-Var does not enforce the one-to-one correspondence constraint, *i.e.*, two different features of the probe image may be matched to the same feature of one gallery image. The CMU PIE database is used for comparison. The top-1 recognition accuracy 86.5% from NBNN-Var is worse than the 90.4% accuracy reported for our Image-to-Class method in Table 3.1, demonstrating the effectiveness of enforcing one-to-one correspondence for face recognition.

3.3.2 Human Gait Recognition

The human gait recognition experiments are performed on the large USF HumanID database constructed by Sarkar *et al.* [3], which comprises 1870 videos of 122 people walking walking in elliptical paths in front of the camera. Five covariates are designed in the database to test the robustness of the algorithms. For details of the database, please refer to Section 2.2.1.

As shown in [95, 93, 101, 100, 98], a complete sequence is partitioned into several subsequences according to the gait period length. In the prior works [100, 95, 93, 101], for each sequence, the binary silhouette images within one gait cycle are averaged to obtain several gray-level average silhouette images. In this work, we also use the same binary silhouettes and gait period lengths to generate the average silhouettes, based on the following considerations [95, 93, 100, 98]: 1) the average silhouette representation is more robust than a time series based silhouette representation [3], especially for the noisy binary images that result from current object segmentation algorithms; 2) average silhouette images do not depend on the choice of starting stance in the gait cycle, thus gait alignment in the temporal domain [3] is unnecessary. The average silhouettes are gray-level silhouette images, with darker regions more likely to contain parts of the pedestrian's body. From the images, we clearly observe that different people have different average silhouettes. The binary silhouette images in one cycle and the corresponding average silhouette image are shown and discussed in Figure 2.1 of Section 2.2.3.

The size of the average silhouette image is 128×88 pixels. Similar to [101], we extracted 40 Gabor features from five different scales and eight different directions at down-sampled positions. Each gallery image is then represented as $I_{gr} = \{q_1^{gr}, q_2^{gr}, \dots, q_{64 \times 44}^{gr}\}$, where q_j^{gr} represents a 40 dimensional Gabor feature vector extracted at the j -th position, and each probe image is characterized as $I_p = \{p_1, p_2, \dots, p_{315}\}$, where p_i is similarly defined.

We compared our algorithm with several other state-of-the-art algorithms in Tables 3.2 and 3.3, in which Rank-1 indicates that the correct subject is ranked as the top-1 candidate, while Rank-5 indicates that the correct subject is ranked among the top-5 candidates, and *Average* is the recognition accuracy among all the probe sets (*i.e.*, the ratio of correctly recognized persons to the total number of persons in all the probe sets). These algorithms include Baseline [3], IMED [111], IMED+LDA, HMM [119], LDA

[95], LDA+Sync [95], LDA+Fusion [95], Dynamics-normalized gait recognition algorithm (DNGR) [92], Matrix based Marginal Fisher Analysis (MMFA) [100], and General Tensor Discriminant Analysis (GTDA) [101]. The results for IMED and IMED+LDA are from [101]. For GTDA, Tao *et al.*, Tao reported multiple results from the combinations of different feature representations and dimensionality reduction algorithms. We only report the best results from the optimal combination in terms of average recognition accuracy in Table 3.2 and 3.3.

In our Image-to-Class method, we need to decide the optimal parameter D^l (See Eq. 3.2). We tried different settings of 0, $\sqrt{2}$ and $2\sqrt{2}$ for D^l . For $D^l = 0$, a probe feature can only match a feature from one of the gallery images if they are located at the same corresponding position, while the larger D^l s of $\sqrt{2}$ and $2\sqrt{2}$ allow matching if the gallery features were displaced within an 8-neighborhood and 24-neighborhood respectively. We report the best results with the optimal D^l in Tables 3.2 and 3.3. It is still an open problem to decide the optimal parameter D^l , which will be investigated in future work.

From Tables 3.2 and 3.3, we observe that our Image-to-Class method outperforms most of the prior work including Baseline [3], IMED [111], IMED+LDA, HMM [119], LDA [95], LDA+Sync [95], LDA+Fusion [95] and GTDA [101] in terms of the Rank-1 and Rank-5 average recognition accuracy. While our Image-to-Class method is worse than the DNGR algorithm in [92], we observe that the DNGR algorithm uses additional training data (*i.e.*, manually annotated silhouettes) to train a Population Hidden Markov Model (pHMM). However, the manually annotated silhouettes used in the pHMM model learning are not publicly available. We also observe that the newly proposed Image-to-Class algorithm outperforms MMFA [100] in terms of Rank-1 average recognition accuracy. While the Image-to-Class algorithm is slightly worse than MMFA in terms of Rank-5 average recognition accuracy, in MMFA thousands of parameters are empirically tuned to achieve the best results, which is impractical in real scenarios.

3.4 Conclusion

We have proposed a new Image-to-Class distance for comparing a probe image to all gallery images in a particular class. The probe and gallery images are represented as a set of local features. An integer programming problem is formulated when computing the Image-to-Class distance, in which any feature of the probe image can be matched to only one feature from one of the gallery images. A spatial neighborhood constraint is also utilized such that only features that are within a corresponding neighborhood of specific size in the spatial domain are allowed to be considered for feature matching. We performed exhaustive experiments on the CMU PIE, FERET and FRGC face databases, as well as the USF HumanID gait database. The experiments clearly demonstrate the effectiveness of our Image-to-Class distance. Based on the Image-to-Class distance, it is possible to develop new machine learning algorithms to further improve the classification performance, which will be investigated in the future. For human gait recognition, we also plan to investigate how the optimal parameter D^l may be automatically determined, and also how the proposed Image-to-Class distance may be integrated with other recent work such as [92, 101, 100] for improved performance.

Table 3.2: Rank-1 recognition accuracy (%) comparison on the USF HumanID Gait database. For our Image-to-Class, the number in parentheses is the optimal D' .

Probe Set	A	B	C	D	E	F	G	H	I	J	K	L	Average
Probe Size	122	54	54	121	60	121	60	120	60	120	33	33	—
Baseline [3]	73	78	48	32	22	17	17	61	57	36	3	3	40.96
HMM [119]	89	88	68	35	28	15	21	85	80	58	17	15	53.54
IMDE [111]	75	83	65	25	28	19	16	58	60	42	2	9	42.87
IMDE + LDA	88	86	72	29	33	23	32	54	62	52	8	13	48.64
LDA [95]	87	85	76	31	30	18	21	63	59	54	3	6	48.20
LDA + Sync [95]	83	94	61	50	48	22	33	48	52	34	18	12	48.04
LDA + Fusion [95]	91	94	81	51	57	25	29	62	60	57	9	12	55.83
DNGR [92]	85	89	72	57	66	46	41	83	79	52	15	24	62.81
MMFA [100]	89	94	80	44	47	25	33	85	83	60	27	21	59.9
GTDA [101]	91	93	86	32	47	21	32	95	90	68	16	19	60.58
Image-to-Class (This work)	93 (0)	89 (0)	81 ($\sqrt{2}$)	54 ($2\sqrt{2}$)	52 ($2\sqrt{2}$)	32 ($2\sqrt{2}$)	34 ($\sqrt{2}$)	81 (0)	78 (0)	62 (0)	12 (0)	9 (0)	61.19

 Table 3.3: Rank-5 recognition accuracy (%) comparison on the USF HumanID Gait database. For our Image-to-Class, the number in parentheses is the optimal D' .

Probe Set	A	B	C	D	E	F	G	H	I	J	K	L	Average
Probe Size	122	54	54	121	60	121	60	120	60	120	33	33	—
Baseline [3]	88	93	78	66	55	42	38	85	78	62	12	15	64.54
HMM [119]	—	—	—	—	—	—	—	—	—	—	—	—	—
IMDE [111]	91	93	83	52	59	41	38	86	76	76	12	15	65.31
IMDE + LDA	95	95	90	52	63	42	47	86	86	78	21	19	68.60
LDA [95]	92	93	89	58	60	36	43	90	81	79	12	12	67.37
LDA + Sync [95]	92	96	91	68	69	50	55	80	78	69	39	30	70.85
LDA + Fusion [95]	94	96	93	85	79	52	57	89	86	77	24	21	76.18
DNGR [92]	96	94	89	85	81	68	69	96	95	79	46	39	82.05
MMFA [100]	98	98	94	76	76	57	60	95	93	84	48	39	79.9
GTDA [101]	98	99	97	68	68	50	56	95	99	84	40	40	77.58
Image-to-Class (This work)	97 (0)	98 (0)	93 (0)	81 ($2\sqrt{2}$)	74 ($\sqrt{2}$)	59 ($2\sqrt{2}$)	55 ($\sqrt{2}$)	94 (0)	95 (0)	83 (0)	30 (0)	33 (0)	79.17

Chapter 4

Trace Ratio Based Flexible Semi-supervised Discriminant Analysis

In this chapter, we propose a new dimensionality reduction method called Trace Ratio based Flexible Semi-supervised Discriminant Analysis (TR-FSDA) for face recognition.

4.1 Introduction

Dimensionality reduction methods are very popular in face recognition. Many dimensionality reduction methods aim to find a projection matrix W which maximizes $Tr(W^T A W)$ while minimizes $Tr(W^T B W)$, in which A and B can be formulated as graph relationships that represent different types of information of the data [31]. Take LDA [4] as an example, A and B can be regarded as the between-class scatter matrix and the within-class scatter matrix respectively. As shown in [5], this problem can be formulated in the trace ratio form of $\max_W \frac{Tr(W^T A W)}{Tr(W^T B W)}$, to which the closed form solution does not exist. So it is usually converted to the ratio trace form of $\max_W Tr((W^T B W)^{-1}(W^T A W))$, which can be directly solved by generalized eigenvalue decomposition [4, 30]. Recently, some works have been proposed to directly solve the original trace ratio formulation. Guo et al. [23] proposed a method to solve the trace ratio problem by employing a bisection

algorithm, which iteratively approaches the optimal trace ratio value from both lower and upper directions. However, the solution may not be very accurate according to the experiments shown in [24]. In [5], Wang et al. developed a more accurate and effective algorithm for the trace ratio problem by iteratively solving a related trace difference problem with eigenvalue decomposition. The experiments in [5] also demonstrate that Trace Ratio based LDA (referred to as TRDA here) is generally better than the conventional LDA [4] in which the projection matrix is solved by generalized eigenvalue decomposition. Recently, Jia et al. [24] proposed a new method which is generally faster than [5]. It uses the first order Taylor expansion of the previously computed eigenvalues to approximate the new eigenvalues at the next iteration, such that the eigenvalue decomposition only needs to be conducted twice in [24].

It is well-known that it is much easier to collect the unlabeled samples than the labeled samples. Recently, semi-supervised learning algorithms have been developed to effectively utilize a limited number of labeled samples and a large amount of unlabeled samples for real-world applications. Local and Global Consistency (LGC) [40] and Gaussian Fields and Harmonic Functions (GFHF) [41] propagate the labels from the labeled training data to the unlabeled training data based on label fitness (*i.e.*, the predicted labels of the labeled data should be consistent with their original labels) as well as manifold smoothness (*i.e.*, the neighboring samples in high-dimensional space should share similar decision values). However, LGC/GFHF can not handle the unseen test data. Manifold Regularization (MR) [39, 42, 7] respectively extends Regularized Least Squares and SVM into Laplacian Regularized Least Squares (LapRLS) and Laplacian SVM (LapSVM) by adding a geometrical regularizer. The recent study [9] shows that the linear version of the multi-dimensional extension of LapRLS (referred to as LapRLS/L in this work) [7] can be considered as an out-of-sample extension of LGC/GFHF by enforcing the label prediction matrix F to be in the linear subspace spanned by the training data matrix X

(*i.e.*, $F = X^T W + \mathbf{1}b^T$). Based on this observation, FME [9] relaxes the hard constraint $F = X^T W + \mathbf{1}b^T$ in LapRLS/L by introducing a regularization term $\|F - X^T W - \mathbf{1}b^T\|^2$ into the objective function of LapRLS/L. Similar as MR, Semi-supervised Discriminant Analysis (SDA) [6] also introduces a geometrical regularizer into the objective function of LDA by utilizing the manifold structure of both labeled and unlabeled training data.

Observing that the low-dimensional representation F is also constrained by the subspace spanned by the data matrix (*i.e.*, $F = X^T W$) in many dimensionality reduction algorithms (*e.g.*, PCA, LDA, LPP, SDA), similar to FME, we first relax this hard constraint by substituting $X^T W$ by F and adding a residue term $\|F - X^T W\|^2$ in the objective function of SDA. With such relaxation, our method can better cope with the data points sampled from nonlinear manifold. We also formulate our objective function in a trace ratio form in order to obtain a better projection matrix W . Since there is no closed form solution to the proposed objective function, we develop an iterative algorithm to solve our objective function and our algorithm can be proved to converge to the optimum based on the fast Newton-Raphson method. The resultant projection matrix W can be used for dimensionality reduction for any training data or unseen test data before final classification. Extensive experiments on three face databases Umist [19], FERET [1], and FRGC [20] demonstrate that our proposed algorithm outperforms the existing dimensionality reduction methods.

4.2 Related Work

4.2.1 LDA and Trace Ratio Based LDA

Linear Discriminant Analysis (LDA) [4] aims to maximize the between-class scatter and at the same time minimize the within-class scatter. Define the set of all the labeled samples as $X_L = [x_1, x_2, \dots, x_l] \in \mathbb{R}^{\hat{d} \times l}$. $\hat{d}$ is the feature deminsion and l is the number of labeled samples. Let us also denote C as the number of classes and R_c as the index

set of labeled samples from the c -th class. The within-class scatter matrix is defined as $M_w = \sum_{c=1}^C \sum_{i \in R_c} (x_i - \bar{x}_c)(x_i - \bar{x}_c)^T$, where $\bar{x}_c = \frac{1}{l_c} \sum_{i \in R_c} x_i$ is the mean of the samples in the c -th class, with l_c being the number of elements in R_c . Clearly, we have $l = \sum_{c=1}^C l_c$. The between-class scatter matrix is defined as $M_b = \sum_{c=1}^C l_c (\bar{x}_c - \bar{x})(\bar{x}_c - \bar{x})^T$ where $\bar{x} = \frac{1}{l} \sum_{i=1}^l x_i$ is the mean of all labeled samples. We can also define two similarity matrix $S_w \in \mathbb{R}^{l \times l}$ and $S_b \in \mathbb{R}^{l \times l}$ as $S_w(i, j) = 1/l_c$ and $S_b(i, j) = 1/l - 1/l_c$ if x_i and x_j are both from the c -th class, and $S_w(i, j) = 0$ and $S_b(i, j) = 1/l$ otherwise. We also denote the Laplacian matrices of S_w as $L_w = D_w - S_w \in \mathbb{R}^{l \times l}$, where $D_w \in \mathbb{R}^{l \times l}$ is the diagonal matrix with the diagonal elements $D_w(i, i) = \sum_j S_w(i, j)$. The Laplacian matrix L_b of S_b can be similarly defined. Let us also denote another Laplacian matrix $L_t = L_w + L_b = I - \frac{1}{l} \mathbf{1}\mathbf{1}^T \in \mathbb{R}^{l \times l}$, with $\mathbf{1}$ being the column vector of all ones. As shown in [31], the within-class scatter matrix $M_w = X_L L_w X_L^T$, the between-class scatter matrix $M_b = X_L L_b X_L^T$ and the total scatter matrix $M_t = M_w + M_b = X_L L_t X_L^T$. We have the following objective function for the Trace Ratio based LDA (TRDA) [5]:

$$W^* = \operatorname{argmax}_{W^T W = I} \frac{\operatorname{Tr}(W^T M_b W)}{\operatorname{Tr}(W^T M_t W)}, \quad (\text{Eq. 4.1})$$

where $W \in \mathbb{R}^{\hat{d} \times \mathbf{m}}$ ($\mathbf{m} < \hat{d}$) is the projection matrix. There is no closed form solution for the objective function in Eq. 4.1. In LDA [4], the objective function in Eq. 4.1 is converted from the trace ratio form to the following ratio trace form:

$$W^* = \operatorname{argmax}_W \operatorname{Tr} \left((W^T M_t W)^{-1} (W^T M_b W) \right), \quad (\text{Eq. 4.2})$$

where W^* can be directly solved by using generalized eigenvalue decomposition. Recently, several iterative algorithms were proposed to directly solve the objective function in Eq. 4.1. For more details, please refer to [23, 24, 5].

4.2.2 Semi-supervised Discriminant Analysis (SDA)

In semi-supervised learning, we define the set of all the training samples as $X = [x_1, x_2, \dots, x_l, x_{l+1}, \dots, x_{l+u}] \in \mathbb{R}^{\hat{d} \times (l+u)}$, where $\{x_i\}_{i=1}^l$ and $\{x_i\}_{i=l+1}^{l+u}$ are the labeled and unlabeled samples respectively. X is assumed to be normalized to have a mean of $\mathbf{0}^{\hat{d} \times 1}$ after subtracting the mean of all the samples. Define $S \in \mathbb{R}^{(l+u) \times (l+u)}$ as the similarity matrix with S_{ij} representing the similarity between sample i and j , and D as the diagonal matrix with the diagonal elements $D_{ii} = \sum_j S_{ij}$. The Laplacian matrix $L \in \mathbb{R}^{(l+u) \times (l+u)}$ is defined as $L = D - S$ and the normalized Laplacian matrix is defined as $\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. Semi-supervised Discriminant Analysis (SDA) [6] is a semi-supervised extension of LDA which adds a geometrically based regularization term in the objective function of LDA. The optimal projection matrix W^* can be solved by optimizing the following objective function:

$$W^* = \operatorname{argmax}_W \frac{|W^T X \tilde{L}_b X^T W|}{|W^T (X(\tilde{L}_t + \alpha L) X^T + \beta I) W|}, \quad (\text{Eq. 4.3})$$

where $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$ are the pre-defined parameters that balance different terms while $\tilde{L}_b \in \mathbb{R}^{(l+u) \times (l+u)}$ and $\tilde{L}_t \in \mathbb{R}^{(l+u) \times (l+u)}$ are the augmented matrices of L_b and L_t , namely:

$$\tilde{L}_b = \begin{bmatrix} L_b & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{L}_t = \begin{bmatrix} L_t & 0 \\ 0 & 0 \end{bmatrix}. \quad (\text{Eq. 4.4})$$

Since X is normalized to have a mean of $\mathbf{0}^{\hat{d} \times 1}$, $\tilde{L}_t$ can be replaced by $\tilde{I}$ in [6], where $\tilde{I} \in \mathbb{R}^{(l+u) \times (l+u)}$ is the augmented matrix of the identity matrix. The objective function of SDA is also solved by generalized eigenvalue decomposition.

4.2.3 Local and Global Consistency (LGC) and Gaussian Fields Harmonic Functions (GFHF)

Local and Global Consistency (LGC) [40] and Gaussian Fields Harmonic Functions (GFHF) [41] compute labels for unlabeled data by imposing label fitness and manifold

smoothness constraints. We denote $\{y_i \in \{1, 2, \dots, C\} |_{i=1}^l\}$ as the label set for the labeled data $\{x_i |_{i=1}^l\}$ where C is the number of subjects (classes). We also denote the binary matrix $Y \in \mathbb{B}^{(l+u) \times C}$ with the i -th row denoted as $Y_i. \in \mathbb{B}^{1 \times C}$, representing the label of image i ($i = 1, \dots, l$), namely, $Y_{ij} = 1$ if x_i has a label of $y_i = j$, and $Y_{ij} = 0$ otherwise.

The objective function of LGC is:

$$F^* = \operatorname{argmin}_F \frac{1}{2} \sum_{i,j=1}^{l+u} \left\| \frac{F_{i.}}{\sqrt{D_{ii}}} - \frac{F_{j.}}{\sqrt{D_{jj}}} \right\|^2 S_{ij} + \eta \sum_{i=1}^{l+u} \|F_{i.} - Y_{i.}\|^2, \quad (\text{Eq. 4.5})$$

where $F \in \mathbb{R}^{(l+u) \times C}$ is the prediction label matrix of all the $l + u$ samples, $F_{i.} \in \mathbb{R}^{1 \times C}$ is the i -th row of F , and $\eta \in \mathbb{R}$ is a parameter which balances different terms. GFHF has the objective function as follows:

$$F^* = \operatorname{argmin}_F \frac{1}{2} \sum_{i,j=1}^{l+u} \|F_{i.} - F_{j.}\|^2 S_{ij} + \eta_\infty \sum_{i=1}^l \|F_{i.} - Y_{i.}\|^2, \quad (\text{Eq. 4.6})$$

where $\eta_\infty \in \mathbb{R}$ is a very large number to force the predicted labels of the labeled data to be consistent with their true labels. Eq. 4.5 and Eq. 4.6 can be unified within the objective function as follows [120]:

$$F^* = \operatorname{argmin}_F \operatorname{Tr}(F^T B F) + \operatorname{Tr}(F - Y)^T U (F - Y), \quad (\text{Eq. 4.7})$$

where U is a diagonal matrix with diagonal element $U_{ii} = \eta_l$ when $i = \{1, \dots, l\}$ and $U_{ii} = \eta_u$ when $i = \{l + 1, \dots, l + u\}$. $\eta_l \in \mathbb{R}$ and $\eta_u \in \mathbb{R}$ are the weighting parameters for labeled and unlabeled data respectively. In LGC, we have $B = \tilde{L}$, and $\eta_l = \eta_u = \eta$, while in GFHF, we have $B = L$, $\eta_l = \eta_\infty$ and $\eta_u = 0$. It is obvious that LGC/GFHF can not handle unseen test data.

4.2.4 Linear Laplacian Regularized Least Squares (LapRLS/L)

Linear Laplacian Regularized Least Squares (LapRLS/L) [7] has the following objective function:

$$\begin{aligned} (W^*, \mathbf{b}^*) &= \underset{W, \mathbf{b}}{\operatorname{argmin}} G_1(W, \mathbf{b}) = \underset{W, \mathbf{b}}{\operatorname{argmin}} \left(\gamma_I \operatorname{Tr}(W^T X L X^T W) + \gamma_A \|W\|^2 \right. \\ &\quad \left. + \frac{1}{l} \sum_{i=1}^l \|Y_{i.}^T - W^T x_i - \mathbf{b}\|^2 \right), \end{aligned} \quad (\text{Eq. 4.8})$$

where $\gamma_A \in \mathbb{R}$ and $\gamma_I \in \mathbb{R}$ are two parameters to balance different terms. LapRLS/L is a varied out-of-sample extension of LGC/GFHF. Let us assume $F = X^T W + \mathbf{1}\mathbf{b}^T$, where $\mathbf{1}^{(l+u) \times 1}$ is a column vector with all elements being one and $\mathbf{b} \in \mathbb{R}^{C \times 1}$ is the bias term. Substituting F by $X^T W + \mathbf{1}\mathbf{b}^T$ and letting $B = L$ in Eq. 4.7, we have:

$$\begin{aligned} (W^*, \mathbf{b}^*) &= \underset{W, \mathbf{b}}{\operatorname{argmin}} \left(\operatorname{Tr} \left((X^T W + \mathbf{1}\mathbf{b}^T)^T L (X^T W + \mathbf{1}\mathbf{b}^T) \right) \right. \\ &\quad \left. + \operatorname{Tr} \left((X^T W + \mathbf{1}\mathbf{b}^T - Y)^T U (X^T W + \mathbf{1}\mathbf{b}^T - Y) \right) \right). \end{aligned} \quad (\text{Eq. 4.9})$$

Since L is a Laplacian matrix, we have $L\mathbf{1} = \mathbf{0}$ and $\mathbf{1}^T L = \mathbf{0}^T$, so Eq. 4.9 becomes:

$$(W^*, \mathbf{b}^*) = \underset{W, \mathbf{b}}{\operatorname{argmin}} \left(\operatorname{Tr}(W^T X L X^T W) + \operatorname{Tr}((X^T W + \mathbf{1}\mathbf{b}^T - Y)^T U (X^T W + \mathbf{1}\mathbf{b}^T - Y)) \right). \quad (\text{Eq. 4.10})$$

Adding a regularization term $\frac{\gamma_A}{\gamma_I} \|W\|^2$, setting $\eta_l = \frac{1}{l\gamma_I}$ and $\eta_u = 0$ in Eq. 4.10, we have:

$$\begin{aligned} (W^*, \mathbf{b}^*) &= \underset{W, \mathbf{b}}{\operatorname{argmin}} G_2(W, \mathbf{b}) = \underset{W, \mathbf{b}}{\operatorname{argmin}} \left(\operatorname{Tr}(W^T X L X^T W) \right. \\ &\quad \left. + \frac{\gamma_A}{\gamma_I} \|W\|^2 + \frac{1}{l\gamma_I} \sum_{i=1}^l \|Y_{i.}^T - W^T x_i - \mathbf{b}\|^2 \right). \end{aligned} \quad (\text{Eq. 4.11})$$

It can be observed that $G_2(W, \mathbf{b}) = G_1(W, \mathbf{b})/\gamma_I$, and obviously Eq. 4.8 and Eq. 4.11 have the same solution.

From Proposition 1 we observe that in LapRLS/L, the prediction label matrix F should be within the linear subspace spanned by the data X and the bias $\mathbf{b}$, so LapRLS/L can be treated as a varied out-of-sample extension of LGC/GFHF.

4.2.5 Flexible Manifold Embedding (FME)

Flexible Manifold Embedding (FME) [9] is a general extension of LGC/GFHF and LapRLS/L, which has the following objective function:

$$\begin{aligned}
 (F^*, W^*, \mathbf{b}^*) = & \underset{F, W, \mathbf{b}}{\operatorname{argmin}} Tr(F^T L F) + Tr((F - Y)^T U (F - Y)) + \beta \|W\|^2 \\
 & + \gamma \|F - X^T W - \mathbf{1}\mathbf{b}^T\|^2,
 \end{aligned} \tag{Eq. 4.12}$$

where $\beta \in [0, +\infty)$ and $\gamma \in [0, +\infty)$ are two parameters to balance different terms, and $\eta_l = 1$ and $\eta_u = 0$ are used as the diagonal elements of the matrix U in [9]. LGC/GFHF and LapRLS/L can be seen as the special cases of FME [9]. When $\beta = 0$ and $\gamma = 0$, FME reduces to LGC/GFHF. When γ approaches $+\infty$, FME can reduce to LapRLS/L [9]. In practice, FME relaxes the hard constraint $F = X^T W + \mathbf{1}\mathbf{b}^T$ in LapRLS/L by introducing a residue term to model the mismatch between F and $X^T W + \mathbf{1}\mathbf{b}^T$. The matrix F^* can be used as the label prediction matrix for the unlabeled training data. Additionally, the resultant W^* and $\mathbf{b}^*$ from Eq. 4.12 can be used to cope with the unseen test data.

4.3 Our Proposed Algorithm

4.3.1 Motivation

In many dimensionality reduction methods like PCA, LDA, SDA and LapRLS/L, the low-dimensional representation of the data or the label prediction matrix¹ F is constrained as $F = X^T W$. In order to better cope with the data sampled from nonlinear manifold, similarly as in FME, we also relax the hard constraint $F = X^T W$ in SDA by substituting $X^T W$ by F and adding a residue term $\|F - X^T W\|^2$. Moreover, we further formulate the objective function in a trace ratio form in order to obtain a better projection matrix

¹We ignore the bias term in LapRLS/L in some places for better presentation.

W . Our algorithm Trace Ratio based Flexible Semi-supervised Discriminant Analysis (referred to as TR-FSDA) also solves for the low-dimensional embedding F of training samples and the projection matrix W which can provide low-dimensional mappings for both unlabeled training data and unseen test data.

4.3.2 Our Objective Function and Solution

Our objective function is defined as follows:

$$(F^*, W^*) = \operatorname{argmax}_{F, W^T W = I} \frac{\operatorname{Tr}(F^T \tilde{L}_b F)}{\operatorname{Tr}(F^T \tilde{L}_a F) + \|X^T W - F\|^2 + \beta \operatorname{Tr}(W^T W)}, \quad (\text{Eq. 4.13})$$

where $\tilde{L}_a = \tilde{L}_t + \alpha L$, $\alpha = \mathbb{R}^+$ and $\beta = \mathbb{R}^+$ are two parameters to balance different terms, $W \in \mathbb{R}^{\hat{d} \times \mathbf{m}}$ is the projection matrix and $F \in \mathbb{R}^{(l+u) \times \mathbf{m}}$ is the low-dimensional embedding of all the training samples. To the best of our knowledge, there is no closed-form solution for the objective function in Eq. 4.13, so we propose an iterative procedure in Algorithm 4.1.

Algorithm 4.1 Trace Ratio based Flexible Semi-supervised Discriminant Analysis (TR-FSDA) Algorithm

Initialize $W_1 \in \mathbb{R}^{\hat{d} \times \mathbf{m}}$ as a columnly orthogonal matrix and $F_1 = X^T W_1$. In our experiments, we define $W_1 = \hat{I}$, where $\hat{I}$ a matrix with the elements $\hat{I}_{ii} = 1$, for $i = \{1, \dots, \mathbf{m}\}$, and 0 otherwise.

Compute

$$\lambda_t = \frac{\operatorname{Tr}(F_t^T \tilde{L}_b F_t)}{\operatorname{Tr}(F_t^T \tilde{L}_a F_t) + \|X^T W_t - F_t\|^2 + \beta \operatorname{Tr}(W_t^T W_t)}. \quad (\text{Eq. 4.14})$$

Update W_{t+1} as the eigenvectors corresponding to the $\mathbf{m}$ largest eigenvalues of $X(\lambda_t^2(\lambda_t I + \lambda_t \tilde{L}_a - \tilde{L}_b)^{-1} - \lambda_t I)X^T$.

Update $F_{t+1} = \lambda_t(\lambda_t I + \lambda_t \tilde{L}_a - \tilde{L}_b)^{-1} X^T W_{t+1}$.

Repeat Steps 2, 3 and 4 until $\lambda_{t+1} - \lambda_t < \xi$ (we set $\xi = 10^{-3}$ in this work).

Denote the resultant λ and W at convergence as λ^* and W^* .

The output W^* at convergence is the projection matrix for mapping the data from high-dimensional space to low-dimensional space.

First, we will explain our iterative approach listed in Algorithm 4.1. Next, we will prove that the resultant λ^* from Algorithm 4.1 is the optimum of our objective function in Eq. 4.13. With λ_t from the t -th iteration (see Eq. 4.14), F_{t+1} and W_{t+1} are computed by maximizing the following trace difference problem:

$$\begin{aligned}
 (F_{t+1}, W_{t+1}) &= \operatorname{argmax}_{F, W^T W = I} Q_{\lambda_t}(F, W) \\
 &= \operatorname{argmax}_{F, W^T W = I} \operatorname{Tr}(F^T \tilde{L}_b F) - \lambda_t \operatorname{Tr}(F^T \tilde{L}_a F) - \lambda_t \|X^T W - F\|^2 - \lambda_t \beta \operatorname{Tr}(W^T W) \\
 &= \operatorname{argmax}_{F, W^T W = I} \operatorname{Tr}(F^T (\tilde{L}_b - \lambda_t \tilde{L}_a - \lambda_t I) F) + 2\lambda_t \operatorname{Tr}(W^T X F) - \lambda_t \operatorname{Tr}(W^T X X^T W) \\
 &\quad - \lambda_t \beta \operatorname{Tr}(W^T W). \tag{Eq. 4.15}
 \end{aligned}$$

If $(\tilde{L}_b - \lambda_t \tilde{L}_a - \lambda_t I)$ is negative-definite, $Q_{\lambda_t}(F, W)$ is a concave quadratic function with respect to F , and the maximum value of $Q_{\lambda_t}(F, W)$ can be obtained by finding the zero point of the partial derivative of $Q_{\lambda_t}(F, W)$ with respect to F . Taking the derivative of $Q_{\lambda_t}(F, W)$ with respect to F and setting it to zero, we have:

$$\begin{aligned}
 \frac{\partial Q_{\lambda_t}(F, W)}{\partial F} &= 2 \left(\tilde{L}_b F - \lambda_t \tilde{L}_a F - \lambda_t (F - X^T W) \right) = 0 \\
 \Rightarrow F &= \lambda_t (\lambda_t I + \lambda_t \tilde{L}_a - \tilde{L}_b)^{-1} X^T W = \lambda_t Z_t X^T W, \tag{Eq. 4.16}
 \end{aligned}$$

where $Z_t = (\lambda_t I + \lambda_t \tilde{L}_a - \tilde{L}_b)^{-1}$. In most cases, $(\tilde{L}_b - \lambda_t \tilde{L}_a - \lambda_t I)$ is negative-definite in our experiments, which means $(\lambda_t I + \lambda_t \tilde{L}_a - \tilde{L}_b)$ is positive definite, hence Z_t is also positive-definite. Additionally, Z_t is also symmetric (*i.e.*, $Z_t = Z_t^T$) because I , $\tilde{L}_a$ and $\tilde{L}_b$ are all symmetric matrices. Substituting Eq. 4.16 in Eq. 4.15, we have:

$$\begin{aligned}
 W_{t+1} &= \operatorname{argmax}_{W^T W = I} \left(\lambda_t^2 \operatorname{Tr} \left((Z_t X^T W)^T (\tilde{L}_b - \lambda_t \tilde{L}_a) (Z_t X^T W) \right) \right. \\
 &\quad \left. - \lambda_t \|X^T W - \lambda_t Z_t X^T W\|^2 - \lambda_t \beta \operatorname{Tr}(W^T W) \right) \\
 &= \operatorname{argmax}_{W^T W = I} \operatorname{Tr} \left(W^T X (\lambda_t^2 Z_t - \lambda_t I) X^T W \right) - \lambda_t \beta \operatorname{Tr}(W^T W) \\
 &= \operatorname{argmax}_{W^T W = I} \operatorname{Tr} \left(W^T X (\lambda_t^2 (\lambda_t I + \lambda_t \tilde{L}_a - \tilde{L}_b)^{-1} - \lambda_t I) X^T W \right). \tag{Eq. 4.17}
 \end{aligned}$$

In Eq. 4.17, we use the property $\tilde{L}_b - \lambda_t \tilde{L}_a = \lambda_t I - Z_t^{-1}$ in the first step. We also have $Tr(W^T W) = \mathbf{m}$ because $W^T W = I$ in the second step. Hence, W_{t+1} is composed of the eigenvectors corresponding to the $\mathbf{m}$ largest eigenvalues of $X(\lambda_t^2(\lambda_t I + \lambda_t \tilde{L}_a - \tilde{L}_b)^{-1} - \lambda_t I)X^T$ (See Step 3 in Algorithm 4.1). Then we can update $F_{t+1} = \lambda_t Z_t X^T W_{t+1}$ (See Step 4 in Algorithm 4.1). Based on W_{t+1} and F_{t+1} , we substitute $F_{t+1} = \lambda_t Z_t X^T W_{t+1}$ back to Eq. 4.14 and we arrive at:

$$\begin{aligned} \lambda_{t+1} &= \frac{Tr\left((\lambda_t Z_t X^T W_{t+1})^T \tilde{L}_b \lambda_t Z_t X^T W_{t+1}\right)}{Tr\left((\lambda_t (Z_t X^T W_{t+1})^T \tilde{L}_a \lambda_t (Z_t X^T W_{t+1})) + \|X^T W_{t+1} - \lambda_t Z_t X^T W_{t+1}\|^2 + \beta Tr(W_{t+1}^T W_{t+1})\right)} \\ &= \frac{\lambda_t^2 Tr\left(W_{t+1}^T X Z_t \tilde{L}_b Z_t X^T W_{t+1}\right)}{Tr\left(W_{t+1}^T X (\lambda_t Z_t \tilde{L}_b Z_t + I - \lambda_t Z_t) X^T W_{t+1}\right) + \beta Tr(W_{t+1}^T W_{t+1})}. \end{aligned} \quad (\text{Eq. 4.18})$$

Again, in the first step of Eq. 4.18, we use the property $\lambda_t \tilde{L}_a = \tilde{L}_b + Z_t^{-1} - \lambda_t I$. Then we will prove that the value of λ obtained using Algorithm 4.1 monotonically increases and its value at convergence can reach the maximum of our objective function in Eq. 4.13.

Proposition 4.3.1 *The value of λ obtained using Algorithm 4.1 monotonically increases, i.e., $\lambda_{t+1} \geq \lambda_t$.*

Proof: According to Eq. 4.15 - Eq. 4.17, for the given λ_t , we have $(F_{t+1}, W_{t+1}) = \arg\max_{F, W^T W = I} Q_{\lambda_t}(F, W)$. So we have $Q_{\lambda_t}(F_{t+1}, W_{t+1})$ with the optimal F_{t+1} and W_{t+1} is not less than $Q_{\lambda_t}(F_t, W_t)$, namely,

$$\begin{aligned} & Tr\left(F_{t+1}^T (\tilde{L}_b - \lambda_t \tilde{L}_a) F_{t+1}\right) - \lambda_t \|X^T W_{t+1} - F_{t+1}\|^2 - \lambda_t \beta Tr(W_{t+1}^T W_{t+1}) \\ & \geq Tr\left(F_t^T (\tilde{L}_b - \lambda_t \tilde{L}_a) F_t\right) - \lambda_t \|X^T W_t - F_t\|^2 - \lambda_t \beta Tr(W_t^T W_t) = 0. \end{aligned}$$

Then, we arrive at:

$$Tr(F_{t+1}^T \tilde{L}_b F_{t+1}) \geq \lambda_t \left(Tr(F_{t+1}^T \tilde{L}_a F_{t+1}) + \|X^T W_{t+1} - F_{t+1}\|^2 + \beta Tr(W_{t+1}^T W_{t+1}) \right). \quad (\text{Eq. 4.19})$$

Obviously, $\tilde{L}_t$ and L are positive-semidefinite, so $\tilde{L}_a = L_t + \alpha L$ is also positive-semidefinite for any $\alpha > 0$. Hence $Tr(F_{t+1}^T \tilde{L}_a F_{t+1}) \geq 0$. Additionally, we have $\beta > 0$ and $Tr(W_{t+1}^T W_{t+1}) = \mathbf{m} > 0$, so we have $Tr(F_{t+1}^T \tilde{L}_a F_{t+1}) + \|X^T W_{t+1} - F_{t+1}\|^2 + \beta Tr(W_{t+1}^T W_{t+1}) > 0$. Hence, the following inequality holds:

$$\lambda_{t+1} = \frac{Tr(F_{t+1}^T \tilde{L}_b F_{t+1})}{Tr(F_{t+1}^T \tilde{L}_a F_{t+1}) + \|X^T W_{t+1} - F_{t+1}\|^2 + \beta Tr(W_{t+1}^T W_{t+1})} \geq \lambda_t.$$

So λ is monotonically increasing after each iteration.

Proposition 4.3.2 *The iterative process in Algorithm 4.1 converges, and the optimal value of λ after convergence is the maximum λ^* of the objective function in Eq. 4.13.*

Proof: From the definition of the maximum λ^* , we have:

$$\lambda^* = \max_{F, W^T W = I} \frac{Tr(F^T \tilde{L}_b F)}{Tr(F^T \tilde{L}_a F) + \|X^T W - F\|^2 + \beta Tr(W^T W)}.$$

Then, for any F and W , we have: $\frac{Tr(F^T \tilde{L}_b F)}{Tr(F^T \tilde{L}_a F) + \|X^T W - F\|^2 + \beta Tr(W^T W)} \leq \lambda^*$,

$$\text{or } Tr\left(F^T (\tilde{L}_b - \lambda^* \tilde{L}_a) F\right) - \lambda^* \|X^T W - F\|^2 - \lambda^* \beta Tr(W^T W) \leq 0.$$

Note that the maximum value 0 can be obtained with the optimal F^* and W^* at convergence, namely,

$$\max_{F, W^T W = I} Tr\left(F^T (\tilde{L}_b - \lambda^* \tilde{L}_a) F\right) - \lambda^* \|X^T W - F\|^2 - \lambda^* \beta Tr(W^T W) = 0. \quad (\text{Eq. 4.20})$$

If we define the following function with respect to the variable of λ :

$$f(\lambda) = \max_{F, W^T W = I} Tr\left(F^T (\tilde{L}_b - \lambda \tilde{L}_a) F\right) - \lambda \|X^T W - F\|^2 - \lambda \beta Tr(W^T W).$$

Substituting F by $\lambda Z X^T W$ according to Eq. 4.16, we have:

$$f(\lambda) = \max_{W^T W = I} Tr\left(W^T X (\lambda^2 Z - \lambda I) X^T W\right) - \lambda \beta Tr(W^T W). \quad (\text{Eq. 4.21})$$

Obviously, λ^* is the solution of $f(\lambda) = 0$. According to Algorithm 4.1 (also see Eq. 4.17), the function value at the point λ_t becomes:

$$\begin{aligned} f(\lambda_t) &= \max_{W^T W = I} \text{Tr}\left(W^T X(\lambda_t^2 Z_t - \lambda_t I) X^T W\right) - \lambda_t \beta \text{Tr}(W^T W) \\ &= \text{Tr}\left(W_{t+1}^T X(\lambda_t^2 Z_t - \lambda_t I) X^T W_{t+1}\right) - \lambda_t \beta \text{Tr}(W_{t+1}^T W_{t+1}), \quad (\text{Eq. 4.22}) \end{aligned}$$

Taking the derivative of $f(\lambda_t)$ at the point λ_t , we have:

$$\begin{aligned} f'(\lambda_t) &= \text{Tr}\left(W_{t+1}^T X(2\lambda_t Z_t - I - \lambda_t^2 Z_t(I + \tilde{L}_a)Z_t) X^T W_{t+1}\right) - \beta \text{Tr}(W_{t+1}^T W_{t+1}) \\ &= \text{Tr}\left(W_{t+1}^T X(2\lambda_t Z_t - I - \lambda_t^2 Z_t \tilde{L}_a Z_t - \lambda_t^2 Z_t I Z_t) X^T W_{t+1}\right) - \beta \text{Tr}(W_{t+1}^T W_{t+1}) \\ &= \text{Tr}\left(W_{t+1}^T X\left(-\lambda_t^2 Z_t \tilde{L}_a Z_t - (\lambda_t Z_t - I)^T (\lambda_t Z_t - I)\right) X^T W_{t+1}\right) - \beta \mathbf{m}. \end{aligned} \quad (\text{Eq. 4.23})$$

Since $\tilde{L}_a$ is positive-semidefinite and Z_t is positive-definite, $-\lambda_t^2 Z_t \tilde{L}_a Z_t$ is negative-semidefinite.

As $-(\lambda_t Z_t - I)^T (\lambda_t Z_t - I)$ is also negative-semidefinite, and $-\beta \mathbf{m} < 0$ for any $\beta > 0$, we have $f'(\lambda_t) < 0$.

So $f(\lambda)$ is monotonically decreasing at any point. Moreover, it can be easily proved that $f(0) \geq 0$ and $f(+\infty) < 0$, and $f(\lambda)$ is continuous. So, the solution of $f(\lambda) = 0$ uniquely exists, hence λ^* exists. Since λ is monotonically increasing after each iteration of our Algorithm 4.1, and the maximum value of λ exists, so the iterative process in Algorithm 4.1 converges, and the value of λ at convergence is the maximum of the objective function.

Proposition 4.3.3 *The iterative process in Algorithm 4.1 is a Newton-Raphson method, thus it converges very fast.*

Proof: From the definition of Newton-Raphson method [121], we can represent λ_{t+1} by:

$$\begin{aligned}
 \lambda_{t+1} &= \lambda_t - \frac{f(\lambda_t)}{f'(\lambda_t)} \\
 &= \lambda_t - \frac{\text{Tr}\left(W_{t+1}^T X(\lambda_t^2 Z_t - \lambda_t I) X^T W_{t+1}\right) - \lambda_t \beta \text{Tr}(W_{t+1}^T W_{t+1})}{\text{Tr}\left(W_{t+1}^T X(2\lambda_t Z_t - I - \lambda_t^2 Z_t(I + \tilde{L}_a)Z_t) X^T W_{t+1}\right) - \beta \text{Tr}(W_{t+1}^T W_{t+1})} \\
 &= \frac{\lambda_t^2 \text{Tr}\left(W_{t+1}^T X Z_t \tilde{L}_b Z_t X^T W_{t+1}\right)}{\text{Tr}\left(W_{t+1}^T X(\lambda_t Z_t \tilde{L}_b Z_t + I - \lambda_t Z_t) X^T W_{t+1}\right) + \beta \text{Tr}(W_{t+1}^T W_{t+1})}. \quad (\text{Eq. 4.24})
 \end{aligned}$$

In Eq. 4.24, we also use the property $\tilde{L}_b + Z_t^{-1} = \lambda_t \tilde{L}_a + \lambda_t I$ in the second step. It can be seen that the value of λ_{t+1} in Eq. 4.24 is exactly the same as that in Eq. 4.18 from the iterative process of Algorithm 4.1. Therefore, our iterative algorithm approaches λ^* based on the Newton-Raphson method, which converges very fast to the maximum. When using the Newton-Raphson method, the resultant $\hat{\lambda}$ must satisfy that $f(\hat{\lambda}) = 0$ at convergence, namely:

$$\hat{\lambda} = \hat{\lambda} - \frac{f(\hat{\lambda})}{f'(\hat{\lambda})} \Rightarrow \frac{f(\hat{\lambda})}{f'(\hat{\lambda})} = 0 \Rightarrow f(\hat{\lambda}) = 0. \quad (\text{Eq. 4.25})$$

From Eq. 4.25 and the proof of Proposition 4.3.2, we have $\hat{\lambda} = \lambda^*$, namely our iterative process in Algorithm 4.1 converges to the maximum λ^* based on the Newton-Raphson method.

4.3.3 Discussion With the Existing Methods

While TR-FSDA and FME [9] both relax the hard constraint $F = X^T W$ as a flexible residue term $\|F - X^T W\|^2$, TR-FSDA is intrinsically different from FME. FME is motivated from the label propagation methods, in which F is the label prediction matrix of the training samples. In contrast, TR-FSDA is based on the discriminant analysis techniques

which have been successfully used for face recognition and other pattern recognition tasks by maximizing the between-class scatter and at the same time minimizing the within-class scatter. Moreover, in TR-FSDA, F is the low-dimensional representation of the training samples and the objective function in TR-FSDA can be readily formulated in a trace ratio form, which can lead to better recognition accuracies.

TR-FSDA is also different from the existing trace ratio based algorithms (*e.g.*, TRDA [5]). First, those algorithms are supervised learning algorithms while TR-FSDA is a semi-supervised learning method. Second, TR-FSDA further relaxes the hard constraint $F = X^T W$ by introducing a flexible regularizer $\|F - X^T W\|^2$ such that TR-FSDA can better cope with the data points sampled from nonlinear manifold.

4.4 Experiments

We test our TR-FSDA on three face databases Umist [19], FERET [1], and FRGC [20]. The sample images from the three databases are displayed in Figure 4.1. We compare TR-FSDA with the semi-supervised dimensionality reduction methods SDA [6], LapRLS/L [7], Transductive Component Analysis (TCA) [8] and FME [9] as well as supervised methods LDA [4] and Trace Ratio based LDA (TRDA) [5]. Note that only the labeled training samples are used for learning the projection matrix in LDA and TRDA.

Similar as in [9], we first use PCA to reduce the dimension of all the training data by preserving 95% energy. In SDA, TCA and TR-FSDA, there are two predefined parameters α and β . In this work, we set them both in the range of $\{10^{-6}, 10^{-3}, 10^0, 10^3, 10^6\}$ and report the best recognition results from the optimal parameter combination. We observe that Z_t in our TR-FSDA is positive-definite in most cases under this parameter setting, which ensures the convergence of our method. For TR-FSDA, we ignore some parameter combinations in which Z_t is not positive-definite, and report the best results from the remaining parameter combinations. In LapRLS/L (*resp.* FME), we also set the

Figure 4.1: Sample images cropped from three face databases Umist (first row), FERET (second row) and FRGC (third row).

two parameters γ_I and γ_A (*resp.* β and γ) in the range of $\{10^{-6}, 10^{-3}, 10^0, 10^3, 10^6\}$ and report the best recognition results. For LapRLS/L and FME, the dimension value $\mathbf{m}$ of the projection matrix W is fixed as C . For LDA, TRDA, SDA, TCA, and TR-FSDA, we need to determine the optimal dimension value $\mathbf{m}$. For TRDA, TCA, and TR-FSDA, we traverse the values of $\mathbf{m}$ in the range of $[\frac{C}{5}, 2C]$ by a step size of $\lfloor (2C - \frac{C}{5})/10 \rfloor$ and report the best recognition results with the optimal dimension values. Note that for the ratio trace based algorithms LDA and SDA, there are no more than $C - 1$ discriminant dimensions because the rank of L_b is bounded at $C - 1$. For LDA and SDA, we therefore only report the top-1 recognition rates by traversing $\mathbf{m}$ in the range of $[\frac{C}{5}, C - 1]$ with the same step size. In contrast, there are more than $C - 1$ discriminant directions in our TR-FSDA because the formulation of our objective function is in the trace ratio form. After projecting all the data to the low-dimensional spaces, we use Nearest Neighbor Classifier [18] for final classification.

Umist [19] database contains 575 images from 20 subjects taken under different view angles. In our experiments, the images are resized to 23×28 pixels. Half of the images per subject are randomly selected as the training data and the rest images are used as the test data. Among the training data, we also randomly choose two images per subject as the labeled training data and employ the remaining images as the unlabeled training data.

We use a subset of the FERET [1] database which consists of 1400 images of 200 subjects. There are seven images per subject taken under seven different views with the angles in the range $\{-40^\circ, -25^\circ, -15^\circ, 0^\circ, +15^\circ, +25^\circ, +40^\circ\}$, which is the same as the database used in Section 3.3.1. All the images are resized to 36×36 . In this work, four images per subject are randomly selected as the training data and the remaining three images are used as the test data. From the four training samples, we randomly select two images as the labeled samples and the other two images are employed as unlabeled training data.

We also use the FRGC1.0 database [20] in our experiments. As large variations in illumination and facial expressions as well as the presence of glasses and blur can be observed in FRGC database, it is a challenging database for face recognition. All images are resized to 36×36 pixels. Since it is a large database, we only choose 20% of the images as the training data and use the remaining 80% images as the test data. Two samples per subject are randomly selected from the training data as the labeled samples, and the rest training images are used as unlabeled training samples. Details of the above face databases can also be found in Section 2.1.1.

For fair comparison, we perform ten rounds of random partitions of training and test data and report the mean recognition rates and standard deviations on the three databases in Table 4.1, in which the top-1 recognition rates on unlabeled training data and unseen test data are denoted by 'U' and 'T' respectively. From Table 4.1, we have the following observations.

First, among the existing semi-supervised learning algorithms including SDA, LapRLS/L, TCA and FME, there is no consistent winner on all the three databases. In terms of the mean recognition rates over ten rounds, SDA achieves the best performance on the FERET and FRGC databases while FME is the best on the Umist among the four algorithms. FME is also consistently better than LapRLS/L in all the experiments. This

Table 4.1: Top-one recognition rates (mean $\pm$ standard deviation%) of LDA [4], TRDA [5], SDA [6], LapRLS/L [7], TCA [8], FME [9] and our TR-FSDA on the three databases. The results shown in boldface are significantly better than the others, judged by the t-test (with a significance level of 0.05).

Algorithms	LDA	TRDA	SDA	LapRLS/L	TCA	FME	TR-FSDA
Umist							
U	62.94 $\pm$ 4.16	68.73 $\pm$ 5.63	77.14 $\pm$ 5.74	76.65 $\pm$ 5.38	75.63 $\pm$ 3.29	77.76 $\pm$ 6.69	81.31 $\pm$ 4.23
T	63.97 $\pm$ 4.22	70.00 $\pm$ 4.64	77.28 $\pm$ 4.15	77.10 $\pm$ 4.39	76.48 $\pm$ 3.47	77.59 $\pm$ 5.22	82.24 $\pm$ 3.48
FERET							
U	65.98 $\pm$ 7.43	76.90 $\pm$ 2.94	70.53 $\pm$ 2.43	51.70 $\pm$ 2.66	60.98 $\pm$ 2.64	52.93 $\pm$ 2.59	79.50 $\pm$ 2.59
T	64.88 $\pm$ 6.43	76.22 $\pm$ 1.46	69.58 $\pm$ 2.35	51.00 $\pm$ 1.53	59.52 $\pm$ 2.06	52.22 $\pm$ 1.54	78.33 $\pm$ 1.46
FRGC							
U	54.94 $\pm$ 1.55	59.61 $\pm$ 1.14	60.50 $\pm$ 1.30	59.19 $\pm$ 1.37	57.40 $\pm$ 1.48	59.29 $\pm$ 1.47	62.01 $\pm$ 1.18
T	53.70 $\pm$ 0.71	57.69 $\pm$ 0.93	58.38 $\pm$ 1.02	57.21 $\pm$ 0.95	55.61 $\pm$ 1.03	57.34 $\pm$ 1.05	59.93 $\pm$ 1.03

Figure 4.2: Trace ratio values (*i.e.*, λ) of our objective function after different number of iterations on three databases. Our algorithm usually converges to the optimal values after less than eight iterations.

is consistent with [9] because LapRLS/L is the special case of FME [9]. Note that on FERET database, the performance of LapRLS/L is not good, probably because of the strong pose variation. In this case, the performance of FME is also not very good because the formulation of FME is based on LapRLS/L.

Second, in most cases, the semi-supervised learning algorithms SDA, LapRLS/L, TCA and FME outperform ratio traced based LDA [4] (*i.e.*, LDA in Table 4.1), owing to the utilization of the unlabeled training data. TRDA is much better than LDA in all the experiments in terms of the mean recognition rates, which is consistent with the previous work [5]. Moreover, TRDA is even better than SDA, LapRLS/L, TCA and FME on the FERET database, demonstrating the effectiveness of the trace ratio method for face recognition.

Finally, our TR-FSDA algorithm achieves the best results on all the three databases in terms of the mean recognition rates on both unlabeled training data and unseen test data. TR-FSDA is also significantly better than the other methods, judged by t-test with a significance level of 0.05. Moreover, as shown in Figure 4.2, our TR-FSDA algorithm usually converges to the optimal values after less than eight iterations on all the three databases because it computes the optimum based on the Newton-Raphson method.

4.5 Conclusion

In this chapter, we have proposed the Trace Ratio based Flexible Semi-supervised Discriminant Analysis (TR-FSDA) algorithm for semi-supervised dimensionality reduction. First, we relax the hard constraint $F = X^T W$ in the conventional dimensionality reduction methods PCA, LDA and SDA in order to better cope with the data points sampled from nonlinear manifold. Specifically, we propose a new objective function by substituting $X^T W$ by F and introducing a residue term $\|F - X^T W\|^2$ into the objective function of SDA. We further formulate the objective function in a trace ratio form and propose an iterative solution to solve the objective function. We also prove that our iterative algorithm is a Newton-Raphson method such that it can reach the maximum of the objective function very fast. We compare TR-FSDA with the existing dimensionality reduction algorithms using three face databases and the extensive experiments demonstrate the effectiveness of our method.

Chapter 5

Patch Distribution Feature Based Methods for Face and Human Gait Recognition

In this chapter, we first introduce the Patch Distribution Feature (PDF) proposed by Lucey and Chen [55], which is proved to be useful for face recognition. Based on PDF, we propose two different dimensionality reduction methods for face and human gait recognition in supervised and semi-supervised settings respectively. For human gait recognition, we propose the Regularized Trace Ratio Discriminant Analysis (RTRDA) based on PDF in supervised setting. Then we introduce the Patch Distribution Compatible Semi-Supervised Dimensionality Reduction (PDC-SSDR) for both face and human gait recognition in semi-supervised setting.

5.1 Patch Distribution Features

In the literature of face recognition and human gait recognition, a face image or a gait silhouette image is usually represented by either a holistic feature vector of very high dimensions [22, 4, 30] or a 2-D or 3-D tensor [100, 101, 45], then dimensionality reduction methods are developed for mapping data from high-dimensional spaces to low-dimensional spaces.

Different from these holistic features, local patch features are also proposed. As shown in [55], an image can be represented by a set of local patch features extracted at selected positions of the image, then patch distributions can be further used as the image features. The Gaussian Mixture Model (GMM) is a well-known model for characterizing distributions, thus it is used to learn the Patch Distribution Feature. Since face and average gait silhouette images are pre-aligned before recognition, we adopt the X-Y patch definition proposed in [57] by concatenating the X-Y coordinates into the local patch features.

Define an image $I = \{p_h\}_{h=1}^H$, where $p_h \in \mathbb{R}^{d \times 1}$ is the local patch feature extracted at pixel h , d is the feature dimension, and H is the number of patches extracted from the image. Also, $p_h = [q_h^T, X_h, Y_h]^T$, where $X_h \in \mathbb{R}, Y_h \in \mathbb{R}$ are the X and Y coordinates of pixel h , $q_h \in \mathbb{R}^{\omega \times 1}$ is the local appearance feature extracted at position (X_h, Y_h) , and thus $d = \omega + 2$.

5.1.1 Local Patch Feature Extraction

There are several choices for extracting the local appearance feature q_h at each pixel. In Chapter 3, for face recognition, the gray-level values in each window centered at pixel h are first normalized to zero mean and unit standard deviation, resulting in the *normalized* features. Then in [57], 2-D Discrete Cosine Transform (DCT) is further performed on the *normalized* features and the coefficients are employed as the local appearance feature q_h for pixel h .

In this subsection, we propose to use another feature as the local appearance feature for training the GMM for human gait recognition. Recently, Gabor wavelet feature has been demonstrated to be an effective feature for human gait recognition. In Chapter 3, the proposed method directly calculated the Image-to-Class distance for human gait recognition by representing each average silhouette image as a set of Gabor features,

Figure 5.1: Real part of the Gabor wavelet kernels $\{\psi_{s,o}|_{s=0}^{s=4}|_{o=0}^{o=7}\}$.

and Tao et al. [101] proposed a tensor subspace learning method called General Tensor Discriminant Analysis (GTDA) by representing the Gabor features extracted from each average silhouette image as a higher order tensor. In this work, we propose to use Gabor features as the local patch feature to learn a new Patch Distribution Feature for human gait recognition.

Let us define a family of Gabor kernel function at one given pixel $\vec{x} = (x, y)$ as:

$$\psi_{s,o}(\vec{x}) = \frac{\|\vec{r}_{s,o}\|^2}{\delta^2} \exp\left(-\frac{\|\vec{r}_{s,o}\|^2 \|\vec{x}\|^2}{2\delta^2}\right) [\exp(i\vec{r}_{s,o} \cdot \vec{x}) - \exp(-\frac{\delta^2}{2})], \quad (\text{Eq. 5.1})$$

where $\vec{r}_{s,o} = \theta_s e^{i\phi_o}$ determines the scale and the orientation of the Gabor function, where $\theta_s = 2^{-\frac{s+2}{2}}\pi$ and $\phi_o = \frac{o}{8}\pi$. In this work, we set the parameter δ as 2π , and also set $s = \{0, 1, 2, 3, 4\}$ and $o = \{0, 1, 2, 3, 4, 5, 6, 7\}$. Then, we have 40 Gabor kernel functions from five different scales and eight different orientations. The term $\exp(-\frac{\delta^2}{2})$ is subtracted in Eq. 5.1 in order to make the kernel DC-free such that the extracted Gabor feature is robust to the illumination. Figure 5.1 shows the magnitudes of the 40 Gabor kernel functions. For each average silhouette image with the size $N_1 \times N_2$, we

first acquire 40 Gabor-filtered images after convolving the Gabor kernel functions with images. For computational efficiency, we further down-sample the Gabor filtered images to size $\lfloor \frac{N_1}{2} \rfloor \times \lfloor \frac{N_2}{2} \rfloor$.

Let us denote q_h as the 40-dimensional Gabor feature extracted from 40 Gabor-filtered images at pixel h . Then we also put the coordinate information into the local feature representation. Specifically, we use the augmented Gabor feature $p_h = [q_h^T, X_h, Y_h]^T \in R^{42}$. Finally, we represent the Gabor-filtered images as a set of augmented Gabor features $I = \{p_h\}_{h=1}^H$, where $H = \lfloor \frac{N_1}{2} \rfloor \times \lfloor \frac{N_2}{2} \rfloor$ is the total number of pixels in each down-sampled Gabor-filtered image.

5.1.2 Universal Background Model

After the local patch features $\{p_h\}_{h=1}^H$ from each image are constructed, it is desired to train an image-specific GMM for every gallery or probe image, however, the local features extracted from one image are not enough to directly learn the image-specific GMM. This is because since there are many parameters to be learned (there are $K \times (1 + d + d^2)$ parameters per image, where K is the number of components of the GMM) in each GMM, it is not stable to learn all of them using the local patch features from only one image. Hence, a global GMM (also referred to as the Universal Background Model (UBM) [55]) is first learned by using *all* the local patch features from *all* the training images, and each image-specific GMM is then adapted from the global GMM. As the global GMM is learned from more local patch features, the parameters learned are more accurate, which will also lead to more accurate image-specific GMM, which are adapted from the global GMM.

Define the set of *all* the local patch features from *all* the gallery images as $P = \{p_n\}_{n=1}^N$, where $N = H \times l$, and l is the number of gallery images. The global GMM can be represented as a set of parameters $\theta = \{(m_k, \mu_k, \Sigma_k)\}_{k=1}^K$, where K is the total number

of components in the GMM and $m_k \in \mathbb{R}$, $\mu_k \in \mathbb{R}^{d \times 1}$ and $\Sigma_k \in \mathbb{R}^{d \times d}$ are the weight, the mean vector and the covariance matrix of the k -th Gaussian component, respectively. Note we have the constraint $\sum_{k=1}^K m_k = 1$. The probability density function of the local patch feature p_n can be represented as the weighted sum of K Gaussian probability density functions:

$$f(p_n|\theta) = \sum_{k=1}^K m_k g(p_n; \mu_k, \Sigma_k), \quad (\text{Eq. 5.2})$$

where the probability density function of the k -th Gaussian function is:

$$g(p_n; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(p_n - \mu_k)^T \Sigma_k^{-1} (p_n - \mu_k)}. \quad (\text{Eq. 5.3})$$

We further assume p_n is independent and identically distributed (i.i.d.). Given a set of observations $P = \{p_1, p_2, \dots, p_N\}$ of all gallery images, the likelihood function of the global GMM model can be defined as:

$$\Lambda(P; \theta) = \prod_{n=1}^N \sum_{k=1}^K m_k g(p_n; \mu_k, \Sigma_k). \quad (\text{Eq. 5.4})$$

Taking the logarithm of Eq. 5.4, we have:

$$L(P; \theta) = \sum_{n=1}^N \log \sum_{k=1}^K m_k g(p_n; \mu_k, \Sigma_k). \quad (\text{Eq. 5.5})$$

The maximum likelihood estimate of the model parameters θ in can be solved by Expectation Maximization (EM) algorithm [122], which initializes the parameters and then update them iteratively. K-means clustering is used as the initialization of μ_k and Σ_k , and a uniform weight is used to initialize m_k . At t -th iteration, we have $m_k^{(t)}$, $\mu_k^{(t)}$ and $\Sigma_k^{(t)}$ known, then we can define the following member function of the local patch feature p_n that belongs to the k -th component as:

$$\varphi^{(t)}(k|n) = \frac{m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k=1}^K m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)})}, \quad (\text{Eq. 5.6})$$

and obviously we have $\sum_{k=1}^K \varphi^{(t)}(k|n) = 1$. Then we have the log likelihood at t -th iteration as follows:

$$L(P; \theta^{(t)}) = \sum_{n=1}^N \log \sum_{k=1}^K m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)}). \quad (\text{Eq. 5.7})$$

Using Jensen's inequality [123] on Eq. 5.7, we have:

$$\begin{aligned} L(P; \theta^{(t)}) &= \sum_{n=1}^N \log \sum_{k=1}^K m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)}) \\ &= \sum_{n=1}^N \log \frac{\sum_{k=1}^K \varphi^{(t)}(k|n) \frac{m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)})}{\varphi^{(t)}(k|n)}}{\sum_{k=1}^K \varphi^{(t)}(k|n)} \\ &\geq \sum_{n=1}^N \frac{\sum_{k=1}^K \varphi^{(t)}(k|n) \log \frac{m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)})}{\varphi^{(t)}(k|n)}}{\sum_{k=1}^K \varphi^{(t)}(k|n)} \\ &= \sum_{n=1}^N \sum_{k=1}^K \varphi^{(t)}(k|n) \log \frac{m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)})}{\varphi^{(t)}(k|n)} \\ &= \sum_{n=1}^N \sum_{k=1}^K \varphi^{(t)}(k|n) \log m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)}) \\ &\quad - \sum_{n=1}^N \sum_{k=1}^K \varphi^{(t)}(k|n) \log \varphi^{(t)}(k|n). \end{aligned} \quad (\text{Eq. 5.8})$$

The first and third equality in Eq. 5.8 use the property $\sum_{k=1}^K \varphi^{(t)}(k|n) = 1$, while the inequality is the Jensen's inequality. Since $\varphi^{(t)}(k|n)$ is known in last iteration, the second term of Equation Eq. 5.8 is constant. If we want to maximize $L(P; \theta^{(t)})$, only the following complete log likelihood is to be maximized:

$$\hat{L}(P; \theta^{(t)}) = \sum_{n=1}^N \sum_{k=1}^K \varphi^{(t)}(k|n) \log m_k^{(t)} g(p_n; \mu_k^{(t)}, \Sigma_k^{(t)}). \quad (\text{Eq. 5.9})$$

In the maximization step, since $\mu_k^{(t+1)}, \Sigma_k^{(t+1)}, m_k^{(t+1)}$ are the optimal values of $\mu_k^{(t)}, \Sigma_k^{(t)}, m_k^{(t)}$ that maximize Eq. 5.9, taking the derivatives of the complete log likelihood $\hat{L}$ with respect to the model parameters $\mu_k^{(t)}, \Sigma_k^{(t)}, m_k^{(t)}$ and setting them to zeros, we obtain the values for $\mu_k^{(t+1)}, \Sigma_k^{(t+1)}, m_k^{(t+1)}$ respectively:

$$\begin{aligned}
 \mu_k^{(t+1)} &= \frac{\sum_{n=1}^N \varphi^{(t)}(k|n)p_n}{\sum_{n=1}^N \varphi^{(t)}(k|n)}, \\
 \Sigma_k^{(t+1)} &= \text{diag} \left(\frac{\sum_{n=1}^N \varphi^{(t)}(k|n)(p_n - \mu_k^{(t+1)})(p_n - \mu_k^{(t+1)})^T}{\sum_{n=1}^N \varphi^{(t)}(k|n)} \right), \quad (\text{Eq. 5.10}) \\
 m_k^{(t+1)} &= \frac{\sum_{n=1}^N \varphi^{(t)}(k|n)}{N},
 \end{aligned}$$

where $\text{diag}(\cdot)$ denotes employing the diagonal elements of the matrix $(\cdot)$ as the diagonal values and setting all the off-diagonal elements to zeros. Note that there are many parameters in the full covariance matrix to be estimated, the covariance matrices Σ_k of all Gaussian functions are constrained to be diagonal, which is also demonstrated to be beneficial in the experiments of the previous work [55]. The resultant parameters at convergence are denoted as $\tilde{m}_k$, $\tilde{\mu}_k$, and $\tilde{\Sigma}_k$, which are the parameters of the global GMM.

After the global GMM is trained, any individual gallery or probe image can be represented by an image-specific GMM which is adapted from the global GMM by Maximum a Posteriori (MAP) probability. The global GMM parameters $\tilde{m}_k$, $\tilde{\mu}_k$, and $\tilde{\Sigma}_k$ are used to initialize the image-specific GMM and for simplicity, $\tilde{\Sigma}_k$ remains constant during the adaptation, while $\tilde{m}_k$ and $\tilde{\mu}_k$ are adjusted by the local patch features from the specific image to be adapted (See Section 5.1.3 for more details). The image-specific GMMs are then used for recognition task.

5.1.3 Maximum A Posteriori (MAP) Adaptation

Given the set of local features $\{p_h\}_{h=1}^H$ from one image, and the pre-learned global GMM parameters $\tilde{m}_k$, $\tilde{\mu}_k$, and $\tilde{\Sigma}_k$, each gallery image or probe image can be further represented by an image-specific GMM adapted from the global GMM [55]. Define an intermediate parameter similar to Eq. 5.6 using global GMM parameters $\{(\tilde{m}_k, \tilde{\mu}_k, \tilde{\Sigma}_k)|_{k=1}^K\}$, we have:

$$\varphi(k|h) = \frac{\tilde{m}_k g(p_h; \tilde{\mu}_k, \tilde{\Sigma}_k)}{\sum_{k=1}^K \tilde{m}_k g(p_h; \tilde{\mu}_k, \tilde{\Sigma}_k)}, \quad (\text{Eq. 5.11})$$

which indicates the membership probability of observation p_h belonging to the k -th Gaussian component. Obviously we have $\sum_{k=1}^K \varphi(k|h) = 1$. The k -th component of the image-specific GMM can be adapted as follows:

$$\begin{aligned}\bar{\mu}_k &= \frac{\sum_{h=1}^H \varphi(k|h)p_h}{\sum_{h=1}^H \varphi(k|h)}, \\ \hat{\mu}_k &= (1 - \zeta)\tilde{\mu}_k + \zeta\bar{\mu}_k, \\ \hat{m}_k &= \frac{\sum_{h=1}^H \varphi(k|h)}{H},\end{aligned}\tag{Eq. 5.12}$$

where $\zeta = \frac{n_k}{n_k + r}$ and $n_k = \sum_{h=1}^H \varphi(k|h)$. To enhance the computational efficiency of the adaptation, only the means and weights are adapted as suggested in [55]. $\bar{\mu}_k$ can be explained as the k -th component's expected mean from the observation data I . However, some of $\bar{\mu}_k$ may not be well-learned due to insufficient observation data I . The linear combination of $\bar{\mu}_k$ and $\tilde{\mu}_k$ with weighting parameter ζ is then used to represent the adapted mean $\hat{\mu}_k$ of the k -th component. r is a predefined parameter for the conjugate priors of the distribution of $\hat{\mu}_k$ with respect to $\tilde{\mu}_k$. Every image can be represented by the image-specific GMM parameter set $\{(\hat{m}_k, \hat{\mu}_k, \tilde{\Sigma}_k) |_{k=1}^K\}$, where $\tilde{\Sigma}_k$ is the k -th component's covariance matrix from the global GMM.

For an image i , we denote its k -th adapted mean and weight as $\hat{\mu}_k^i$ and $\hat{m}_k^i$, and following the representation proposed in [124], the Patch Distribution Feature (PDF) of the i -th image, which is expressed as x_i , can be represented by:

$$\begin{aligned}v_k^i &= \left(\sqrt{\frac{\hat{m}_k^i}{2}} \tilde{\Sigma}_k^{-\frac{1}{2}} \hat{\mu}_k^i\right)^T \in \mathbb{R}^{1 \times d}, \\ x_i &= [v_1^i, v_2^i, \dots, v_K^i]^T \in \mathbb{R}^{\hat{d} \times 1},\end{aligned}\tag{Eq. 5.13}$$

where v_k^i is the weighted mean of the k -th image-specific GMM component for image i and $\hat{d} = d \times K$ is the feature dimension of x_i .

5.1.4 Gallery Set Expansion

Nearest Feature Line (NFL) [108] employs virtual training samples generated from any two training images of the same class for face recognition. Inspired by NFL, we also try to construct virtual samples to expand the gallery set. Suppose we have C classes and l_c images for the c -th class in the gallery set, hence we have $l = \sum_{c=1}^C l_c$ gallery images in total. After MAP adaptation for each gallery image, we have l image-specific GMMs. To expand the gallery set, we also train a virtual image-specific GMM by any two gallery images from the same class. The virtual image-specific GMM is also learned using Eq. 5.12 in which $\{p_h|_{h=1}^H\}$ is set as the union of the local patch features from the two images. For the c -th class, there are $\binom{l_c}{2} = \frac{l_c(l_c-1)}{2}$ combinations, so there are $\frac{l_c(l_c-1)}{2}$ virtual image-specific GMMs. We label each virtual GMM according to the label of the two images, and add them to the gallery GMM set, in which case, there are $\bar{l}_c = l_c + \frac{l_c(l_c-1)}{2}$ image-specific GMMs of each class after expansion. Hence, the expanded gallery set consists of $\bar{l} = \sum_{c=1}^C (l_c + \frac{l_c(l_c-1)}{2})$ image-specific GMMs in total. This operation expands the gallery set, making it cover more variation of the gallery images. It is worth mentioning that the X-Y Patch method proposed by [57] does not expand the gallery set and we also conduct experiments to compare the recognition rate with or without gallery set expansion in Section 5.2.4.

5.2 Regularized Trace Ratio Discriminant Analysis (RTRDA) with Patch Distribution Feature for Human Gait Recognition

5.2.1 Introduction

In this section, we adopt the Gabor features to train the Patch Distribution Feature (PDF) by the process proposed in Section 5.1, and every gallery or probe image is then represented as the PDF by Eq. 5.13. Since PDF in Eq. 5.13 is also a high-dimensional

representation, we propose a new dimensionality reduction algorithm called Regularized Trace Ratio Discriminant Analysis (RTRDA), which is a generalized form of Trace Ratio Discriminant Analysis (TRDA), to find the best subspaces to map the data into low-dimensional space for recognition. Experiments on the USF HumanID gait database show that RTRDA in combination with PDF outperforms all the other previous works in terms of average Rank-1 recognition rates.

5.2.2 Proposed Algorithm

Every gallery sample (including virtual sample) or probe sample is now represented by a $\hat{d}$ dimensional PDF x_i and the feature dimension $\hat{d}$ is a large value in gait recognition (2816 in our experiments). Hence, dimensionality reduction methods can be applied to x_i to achieve better performance. Define the set of training data as $X = \{x_i\}_{i=1}^{\bar{l}} \in \mathbb{R}^{\hat{d} \times \bar{l}}$ where $\bar{l}$ is the number of PDF in the expanded gallery set. Linear Discriminant Analysis (LDA) is widely used in face recognition and human gait recognition, which aims to find a subspace to maximize the between-class scatters while minimize the within-class scatters. Similar to Section 4.2, Let us define C as the number of classes and R_c as the index set of training images from the c -th class, The within-class scatter matrix is defined as $M_w = \sum_{c=1}^C \sum_{i \in R_c} (x_i - \bar{x}_c)(x_i - \bar{x}_c)^T \in \mathbb{R}^{\hat{d} \times \hat{d}}$, where $\bar{x}_c = \frac{1}{l_c} \sum_{i \in R_c} x_i$ is the mean of the samples in the c -th class, with l_c being the number of elements in R_c . The between-class scatter matrix is defined as $M_b = \sum_{c=1}^C l_c (\bar{x}_c - \bar{x})(\bar{x}_c - \bar{x})^T$ where $\bar{x} = \frac{1}{\bar{l}} \sum_{i=1}^{\bar{l}} x_i$ is the mean of all samples. The objective of LDA can be formulated as a trace ratio problem:

$$W^* = \underset{\substack{W \in \mathbb{R}^{\hat{d} \times \mathbf{m}} \\ W^T W = I}}{\operatorname{argmax}} \frac{\operatorname{Tr}(W^T M_b W)}{\operatorname{Tr}(W^T M_w W)}. \quad (\text{Eq. 5.14})$$

Where W is the projection matrix and $\mathbf{m} \leq \hat{d}$ (note that it is equivalent to Eq. 4.1). Since there is no direct solution to Eq. 4.1, the problem is usually converted to the following

ratio trace problem:

$$W^* = \operatorname{argmax}_{W \in \mathbb{R}^{\hat{d} \times \mathbf{m}}} \operatorname{Tr} \left((W^T M_w W)^{-1} (W^T M_b W) \right), \quad (\text{Eq. 5.15})$$

which can be solved by generalized eigenvalue decomposition. Specifically in Eq. 5.15, the solution is the eigenvectors corresponding to the $\mathbf{m}$ largest eigenvalues of $M_w^{-1} M_b$ [18]. In [5], an iterative algorithm is proposed to directly solve Eq. 5.14, which is referred to as Trace Ratio Discriminant Analysis (TRDA, also referred to as Trace Ratio based LDA in Chapter 4). TRDA is experimentally shown to be better than LDA in most cases in [5].

In this work, we generalize the TRDA by adding a regularization term in Eq. 5.14 as follows:

$$W^* = \operatorname{argmax}_{\substack{W \in \mathbb{R}^{\hat{d} \times \mathbf{m}} \\ W^T W = I}} \frac{\operatorname{Tr}(W^T M_b W)}{\operatorname{Tr}(W^T (M_w + \beta I) W)}. \quad (\text{Eq. 5.16})$$

where βI is the regularization term with I being the identity matrix and $\beta \in \mathbb{R}^+$ being a predefined parameter. Similar to TRDA, the optimal W^* in Eq. 5.16 can be solved iteratively as Algorithm 5.1 in the following:

Algorithm 5.1 Regularized Trace Ratio Discriminant Analysis Algorithm.

- 1: Initialize $W = \mathbb{R}^{\hat{d} \times \mathbf{m}}$ as a columnly orthogonal matrix. Similar to Section 4.3.2, we also initialize $W = \hat{I}$, where $\hat{I}$ a matrix with the elements $\hat{I}_{ii} = 1$, for $i = \{1, \dots, \mathbf{m}\}$, and 0 otherwise.
 - 2: Compute $\lambda = \operatorname{Tr}(W^T M_b W) / \operatorname{Tr}(W^T (M_w + \beta I) W)$.
 - 3: Construct W as the eigenvectors corresponding to the largest $\mathbf{m}$ eigenvalues of $M_b - \lambda(M_w + \beta I)$.
 - 4: Repeat steps 2 and 3 until W converges.
-

As shown in [5, 24], W converges to its global optimal very fast. After mapping the PDF into lower dimensional subspaces using W^* , we adopt the simple fusion technique in Section 5.2.3 for classification.

5.2.3 Simple Fusion

Since human gait is recognized based on videos, which are represented by a sequence of average silhouette images, we use the classifier proposed in [93] to define the distance between two videos by simple fusion of the distances between pairs of images from different videos. Suppose we have the probe video which is represented by l_q average silhouette images as $\{z_1, \dots, z_{l_q}\}$. Similarly, the gallery video from the c -th class is denoted by $\{x_1, \dots, x_{l_c}\}$. The distance between the probe video and the gallery video from the c -th class can be computed by simple fusion as follows [92]:

$$\mathbf{d}_{ij} = \|z_i - x_j\| \quad (\text{Eq. 5.17})$$

$$\tilde{\mathbf{d}}_c = \text{median}_{i=1}^{l_q} \left(\min_{j=1}^{l_c} d_{ij} \right). \quad (\text{Eq. 5.18})$$

Finally, the probe video can be classified by:

$$c^* = \underset{c}{\operatorname{argmin}} \tilde{\mathbf{d}}_c. \quad (\text{Eq. 5.19})$$

5.2.4 Experiments

Similar to Section 3.3, we also test our algorithm on the largest human gait database, the USF HumanID Database [3]. Details of the database is in Section 2.2.1. Following previous works [95, 125], we also use average silhouette images of every gait cycle in order to better cope with the noisy binary silhouettes and avoid silhouette alignment [93] (please also refer to Section 3.3). Figure 2.1 shows the binary silhouettes and corresponding average silhouettes. There are originally 122 videos (one video per class) of 658 average silhouette images in the gallery set, with 5-6 images per class. After gallery expansion, there are in total 2138 PDF(15 - 21 PDF per class), greatly enriching the gallery set.

To demonstrate the effectiveness of PDF over traditional features, we compare gray-level feature with PDF directly by simple fusion (referred to as Gray and PDF respectively). To show the power of gallery expansion, we also conduct PDF with gallery

Table 5.1: Rank-1 and Rank-5 recognition rate (%) comparison on the USF HumanID Gait database.

Probe Set	A	B	C	D	E	F	G	H	I	J	K	L	Average
Probe Size	122	54	54	121	60	121	60	120	60	120	33	33	–
Rank-1 Result													
Baseline [3]	73	78	48	32	22	17	17	61	57	36	3	3	40.96
HMM [119]	89	88	68	35	28	15	21	85	80	58	17	15	53.54
IMDE + LDA	88	86	72	29	33	23	32	54	62	52	8	13	48.64
LDA + Fusion [95]	91	94	81	51	57	25	29	62	60	57	9	12	55.83
DNGR [92]	85	89	72	57	66	46	41	83	79	52	15	24	62.81
MMFA [100]	89	94	80	44	47	25	33	85	83	60	27	21	59.9
GTDA [101]	91	93	86	32	47	21	32	95	90	68	16	19	60.58
Image-to-Class [125]	93	89	81	54	52	32	34	81	78	62	12	9	61.19
Gray	83	85	67	24	24	9	16	51	47	42	0	6	40.61
Gray-TRDA	86	85	72	25	36	19	19	52	52	51	3	9	45.38
Gray-RTRDA	86	85	72	25	36	19	19	52	52	51	3	9	45.38
PDF [57]	78	85	70	41	36	20	24	69	67	39	3	9	48.27
PDF-EXP	93	87	81	48	55	30	22	81	76	63	6	18	59.61
PDF-EXP-TRDA	96	87	87	58	55	36	41	81	71	69	6	15	63.88
PDF-EXP-RTRDA	95	91	85	59	62	36	36	84	76	73	3	21	65.41
Rank-5 Result													
Baseline [3]	88	93	78	66	55	42	38	85	78	62	12	15	64.54
IMDE + LDA	95	95	90	52	63	42	47	86	86	78	21	19	68.60
LDA + Fusion [95]	94	96	93	85	79	52	57	89	86	77	24	21	76.18
DNGR [92]	96	94	89	85	81	68	69	96	95	79	46	39	82.05
MMFA [100]	98	98	94	76	76	57	60	95	93	84	48	39	79.9
GTDA [101]	98	99	97	68	68	50	56	95	99	84	40	40	77.58
Image-to-Class [125]	97	98	93	81	74	59	55	94	95	83	30	33	79.17
Gray	93	93	91	57	52	31	34	79	79	69	12	15	63.13
Gray-TRDA	93	94	87	58	64	46	52	75	76	73	6	6	66.16
Gray-RTRDA	93	94	87	58	64	47	52	75	76	74	6	9	66.51
PDF [57]	96	93	89	64	67	46	43	88	84	68	15	33	69.72
PDF-EXP	98	94	91	75	72	49	48	95	98	82	24	36	76.46
PDF-EXP-TRDA	98	94	94	78	81	62	59	94	90	85	21	30	79.34
PDF-EXP-RTRDA	98	94	96	79	79	64	60	95	93	86	18	30	80.11

expansion (referred to as PDF-EXP) directly using simple fusion. We compare TRDA and RTRDA on both gray-level feature and the PDF with gallery expansion. We denote Gray-level features with TRDA and RTRDA as Gray-TRDA and Gray-RTRDA respectively. The corresponding methods using PDF with gallery expansion are denoted as PDF-EXP-TRDA and PDF-EXP-RTRDA respectively. We set the number of components K in GMM as 300 for PDF, PDF-EXP, PDF-EXP-TRDA and PDF-EXP-RTRDA. For Gray-TRDA, Gray-RTRDA, PDF-EXP-TRDA and PDF-EXP-RTRDA, PCA is first performed to reduce the feature dimension as suggested in [22]. We try different dimensions for PCA by preserving the energy from 90% to 100% with an interval of 1% and report the best results. For RTRDA, β is set in the range of $[10^{-4}, 10^{-2}, 1, 10^2, 10^4]$ and only the best result is reported. We run all the possible values of $\mathbf{m}$ in Eq. 5.16 for both TRDA and RTRDA.

Table 5.1 shows the results of above mentioned algorithms as well as the best results from all the previous works which are reported in [101]. We also compare with our Image-to-Class distance, which is introduced in Chapter 3. For all algorithms, we only report the best results from all the parameters. We evaluate all the algorithms based on Rank-1 and Rank-5 recognition rates. PDF outperforms Gray in terms of both Rank-1 and Rank-5 recognition rates, demonstrating the effectiveness of GMM-based Patch Distribution Feature. PDF-EXP consistently outperforms PDF, showing that gallery expansion greatly improves the performance. The proposed algorithm PDF-EXP-RTRDA achieves the best result in terms of average Rank-1 recognition rate, and outperforms all the other existing algorithms except DNGR [92] in terms of average Rank-5 recognition rate. It is worth mentioning that although DNGR achieves the best result in terms of Rank-5 recognition rate, it used additional manually cleaned silhouettes for training, which are not publicly available. However, our algorithm is still significantly better than DNGR in terms of the average Rank-1 recognition rate. The best Rank-1 result among all

previous works that use the same training data as this section is from the Image-to-Class method proposed in Chapter 3, which is far less than PDF-EXP-RTRDA. Finally, in most cases, Gray-RTRDA (*resp.* PDF-EXP-RTRDA) outperforms Gray-TRDA (*resp.* PDF-EXP-TRDA), showing the superiority of our proposed RTRDA over TRDA in terms of both average Rank-1 and Rank-5 recognition rates.

5.2.5 Conclusion

In this section, we propose a new supervised dimensionality reduction method called Regularized Trace Ratio Discriminant Analysis (RTRDA) to find the most discriminant subspaces based on the Patch Distribution Feature for human gait recognition. Experiments on USF humanID database show that our proposed algorithm significantly outperforms the existing algorithms and achieves the best recognition results among all the previous works on USF humanID database in terms of average Rank-1 recognition rate.

5.3 Patch Distribution Compatible Semi-Supervised Dimensionality Reduction for Face and Human Gait Recognition

5.3.1 Introduction

In Section 5.2, we have introduced a supervised dimensionality reduction method called Regularized Trace Ratio Discriminant Analysis (RTRDA) based on Patch Distribution Feature (PDF) for human gait recognition. RTRDA is a general dimensionality reduction method which can be used with any feature. In this section, we propose a new semi-supervised dimensionality reduction method which is tailored for Patch Distribution Features for both face and human gait recognition, utilizing the advantages of both holistic and local feature based approaches.

Following Section 5.1, each image (a face image or an average human silhouette image) is encoded as a set of local patch features (*e.g.* DCT feature or Gabor feature) and then, each image is further represented by the PDF trained from the set of local patch features. To specifically cope with the PDF, we develop a new semi-supervised learning algorithm, referred to as Patch Distribution Compatible Semi-Supervised Dimensionality Reduction (PDC-SSDR) for both face and human gait recognition. Similar to the existing semi-supervised dimensionality reduction methods SDA [6] and LapRLS/L [7], we also employ label fitness (*i.e.*, the image-level prediction label matrix F should be close to the given labels for the labeled samples) and manifold smoothness in our objective function. While the Patch Distribution Feature can be directly combined with SDA [6] and LapRLS/L [7], we argue that the assumption that all the labeled and unlabeled samples reside on a manifold in SDA and LapRLS/L may not hold well because the feature dimension of the Patch Distribution Feature is generally very high. Observing that each GMM component generally represents a facial component (*e.g.*, eye, nose or mouth), we believe it is better to assume that the individual GMM components from all the training images reside on a sub-manifold. We introduce a new regularizer based on a set of local sub-manifold smoothness assumptions in our objective function, in which a component-level prediction label matrix F^k is assigned to individual GMM components. The component-level prediction label matrix F^k is also enforced to be consistent with the image-level prediction label matrix F .

As in LapRLS/L [7], we employ a linear regression function $F = X^T W + \mathbf{1b}^T$ to provide embeddings for the unseen test data. Similar to TR-FSDA proposed in Chapter 4, we additionally measure the mismatch between $X^T W + \mathbf{1b}^T$ and F by introducing the regression residual $\|X^T W + \mathbf{1b}^T - F\|^2$ in our objective function. By modeling the regression residual, we show that we can relax the over-strict hard constraint $F = X^T W + \mathbf{1b}^T$ in LapRLS/L such that our method can better cope with the data sampled from

a non-linear manifold. Finally, the optimal solutions of the component-level prediction label matrix F^k , the image-level prediction label matrix F , the projection matrix W and the bias term $\mathbf{b}$ can be simultaneously obtained. Comprehensive experiments on CMU PIE, FERET and AR face databases as well as the USF HumanID gait database clearly demonstrate the effectiveness of our algorithm. The main contributions of this work include:

- We propose a new semi-supervised dimensionality reduction algorithm PDC-SSDR which can effectively cope with the data sampled from a nonlinear manifold by modeling the regression residual and using a sub-manifold assumption tailored for Patch Distribution Features.
- We also apply Patch Distribution Features for human gait recognition. When compared with the state-of-the-art human gait recognition algorithms, we achieve better Rank-1 and Rank-5 recognition accuracy on the benchmark USF HumanID gait database using PDC-SSDR.

5.3.2 Related Work

As shown Eq. 5.13, for an image i , the Patch Distribution Feature x_i is represented as:

$$\begin{aligned} v_k^i &= \left(\sqrt{\frac{\widehat{m}_k^i}{2}} \widetilde{\Sigma}_k^{-\frac{1}{2}} \widehat{\mu}_k^i \right)^T \in \mathbb{R}^{1 \times d} \\ x_i &= [v_1^i, v_2^i, \dots, v_K^i]^T \in \mathbb{R}^{\widehat{d} \times 1}. \end{aligned} \tag{Eq. 5.20}$$

where $\widehat{\mu}_k^i$ and $\widehat{m}_k^i$ are the k -th adapted mean and weight of image i , v_k^i is the weighted mean and $\widehat{d} = d \times K$ is the feature dimension of x_i .

Figure 5.2: Sample images from the CMU PIE database [2] with four GMM components shown in red, green, blue and magenta. Circles of the same color are from GMM components with the same index. Please refer to the PDF file for better viewing.

5.3.3 Patch Distribution Compatible Semi-supervised Dimensionality Reduction (PDC-SSDR)

5.3.3.1 Motivation

In many semi-supervised learning algorithms, one gallery or probe image is represented by a lengthy vector composed of all the pixel values in high-dimensional space. Conventionally, semi-supervised learning algorithms assume that the high-dimensional data reside on a low-dimensional manifold, so the manifold smoothness constraint assumes that two neighboring samples in high-dimensional space have similar low-dimensional embeddings or prediction labels. However, if the images exhibit large variations in pose, illumination and expression, the manifold smoothness assumption may become invalid when using the holistic features. On the other hand, we observe that the set of weighted GMM components have better manifold characteristics.

In this work, we adopt Patch Distribution Feature as described in Section 5.1.3, and represent each image by Eq. 5.20. Define $V_k = \{v_k^i |_{i=1}^{l+u}\}$ as the set of all the k -th weighted means and $\hat{\theta}_k = \{(\hat{\mu}_k^i, \hat{m}_k^i) |_{i=1}^{l+u}\}$ as the set of all the k -th adapted means and weights from all the $l + u$ training samples.

Note that the last two dimensions of the local X-Y patch feature p_h are the X-Y coordinates of pixel h , therefore the last two dimensions of $\hat{\mu}_k^i$ are also the X-Y coordinates of the k -th component center in image i , which can be denoted as $(\hat{X}_k^i, \hat{Y}_k^i)$. An important observation we have is that the set of all the k -th components' coordinates, which is

denoted as $\{(\hat{X}_k^i, \hat{Y}_k^i)_{i=1}^{l+u}\}$, are in similar areas of the face. To illustrate the observation, in Figure 5.2 we display four selected components with indices $k_1, k_2, k_3, k_4 \in \{1, \dots, K\}$ in 12 images which are selected from different subjects of the CMU PIE database [2] with indices $i_1, i_2, \dots, i_{12} \in \{1, \dots, l+u\}$. A circle is drawn at every position $(\hat{X}_{k_\rho}^{i_\rho}, \hat{Y}_{k_\rho}^{i_\rho})$, where $\rho \in \{1, 2, 3, 4\}$ and $\rho \in \{1, 2, \dots, 12\}$ with the circle size proportional to $\hat{m}_{k_\rho}^{i_\rho}$, which represents the importance of the component in that image. To better view the correspondence between different components, we draw circles with indices k_1, k_2, k_3, k_4 in red, green, blue and magenta, respectively.

From Figure 5.2, we can see that all the k_1 -th components (red circles) in all the images fall near the left corner of the nose, while the k_2 -th components (green circles) in all the images are near the cheek under the right eye. The k_3 -th and k_4 -th components (blue and magenta circles) are somewhere near the right chin of each face. Although they have different weights (illustrated by different sizes of circles), the components with the same color are located at similar positions of the face images. Hence, the values in $\{\hat{\mu}_k^i\}_{i=1}^{l+u}$ represent cluster centers of local features extracted near positions $\{(\hat{X}_k^i, \hat{Y}_k^i)_{i=1}^{l+u}\}$, which are in similar areas throughout all the $l+u$ images. We therefore assume that all the elements $v_k^i (i = 1, \dots, l+u)$ in V_k should form a sub-manifold, because v_k^i can be interpreted as the weighted mean local appearance of similar areas of different faces. Hence, instead of using one manifold formed by all the training images, we assume that there are a set of K sub-manifolds constructed by every $V_k (k = 1, \dots, K)$. Additionally, we can regard the elements in V_k as a facial component from different faces (*e.g.*, eye, nose, mouth). It is reasonable to assume different facial components reside on a set of sub-manifolds, so although the face images may not reside on a manifold due to the variations in pose, illumination and expression, GMM components which represent different facial components should be located on different sub-manifolds. We have similar observations on human gait images in terms of the sub-manifold assumption. Moreover,

our experiments in Section 5.3.4 also demonstrate that the sub-manifold assumption based on GMM components works better when compared with the manifold assumption based on holistic features which concatenate all the GMM components into a lengthy feature vector.

Recall there is a hard constraint in LapRLS/L (please refer to Section 4.2.4) that the predicted labels of the samples should be in the subspace spanned by the data X (*i.e.*, $F = X^T W + \mathbf{1b}^T$). It may become over-strict when the samples are from a non-linear manifold. In this work, similar to Chapter 4, we also add a residual term $\|F - X^T W - \mathbf{1b}^T\|^2$ in our objective function to relax the hard constraint such that our method can better cope with the data sampled from a non-linear manifold. The regression function $X^T W + \mathbf{1b}^T$ can also be used to handle unseen test data.

5.3.3.2 Proposed Objective Function

Similarly as Chapter 4, define the sample set of all the training images as $X = [x_1, x_2, \dots, x_l, x_{l+1}, \dots, x_{l+u}] \in \mathbb{R}^{\hat{d} \times (l+u)}$, where $\{x_i\}_{i=1}^l$ and $\{x_i\}_{i=l+1}^{l+u}$ are the labeled and unlabeled samples respectively. We also assume the mean of X is $\mathbf{0}^{\hat{d} \times 1}$ after subtracting the mean of all the samples. We denote $\{y_i \in \{1, 2, \dots, C\}\}_{i=1}^l$ as the label for the data $\{x_i\}_{i=1}^l$ where C is the number of subjects (classes). We also denote the binary matrix $Y \in \mathbb{B}^{(l+u) \times C}$ with the i -th row denoted as $Y_i \in \mathbb{B}^{1 \times C}$, representing the label of image i ($i = 1, \dots, l$), namely, $Y_{ij} = 1$ if x_i has a label of $y_i = j$, and $Y_{ij} = 0$ otherwise. To utilize a set of sub-manifolds formed by each V_k , we define $F^k \in \mathbb{R}^{(l+u) \times C}$ as the component-level prediction label matrix of every item in V_k . We similarly define the i -th row of F^k as $F_{i.}^k$, which is the label of v_k^i . We also define $F \in \mathbb{R}^{(l+u) \times C}$ as the image-level prediction label matrix of all the samples in X , and the i -th row of F , denoted as $F_{i.}$, is the predicted label of x_i . Furthermore, we define $\hat{F}_i = \{F_{i.}^k\}_{k=1}^K$ as the set of predicted labels of all K components from image i , and it is natural to assume that all the elements of $\hat{F}_i$ should be consistent with the image label $F_{i.}$.

The similarity matrix $S^k \in \mathbb{R}^{(l+u) \times (l+u)}$ is defined as $S_{ij}^k = \exp(-\frac{\|v_k^i - v_k^j\|^2}{\sigma_k^2})$ if v_k^i is among the z nearest neighbors of v_k^j and vice versa, and $S_{ij}^k = 0$ otherwise, where $\sigma_k = \sqrt{\frac{-\epsilon}{\log(1/z)}}$ and ϵ is the mean value of squared distances between neighboring pairs [126]. The Laplacian matrix of the k -th component is defined as $L^k = D^k - S^k$, where D^k is the diagonal matrix with the diagonal elements $D_{ii}^k = \sum_j S_{ij}^k$.

We also denote $U \in \mathbb{R}^{(l+u) \times (l+u)}$ as a diagonal matrix with $U_{ii} = 1$ if $i \in \{1, \dots, l\}$ and 0 elsewhere, $W \in \mathbb{R}^{\hat{d} \times C}$ as the dimensionality reduction projection matrix for mapping unseen test data, and $\mathbf{b} \in \mathbb{R}^{C \times 1}$ as the bias term. The objective function of Patch Distribution Compatible Semi-Supervised Dimensionality Reduction (PDC-SSDR) is:

$$\begin{aligned}
 (F^{k*}, F^*, W^*, \mathbf{b}^*) &= \underset{F^k, F, W, \mathbf{b}}{\operatorname{argmin}} \Omega(F^k, F, W, \mathbf{b}) \\
 &= \underset{F^k, F, W, \mathbf{b}}{\operatorname{argmin}} \left(\sum_k \operatorname{Tr}(F^{kT} L^k F^k) \right. \\
 &\quad + \nu \operatorname{Tr}((F - Y)^T U (F - Y)) \\
 &\quad + \tau \sum_k \operatorname{Tr}((F - F^k)^T \Lambda^k (F - F^k)) \\
 &\quad \left. + \gamma \|X^T W + \mathbf{1} \mathbf{b}^T - F\|^2 + \beta \|W\|^2 \right). \quad (\text{Eq. 5.21})
 \end{aligned}$$

where Λ^k is a diagonal matrix with the diagonal element $\Lambda_{ii}^k = \hat{m}_k^i$. In Eq. 5.21, L^k , Y , U , Λ^k and X are known or can be pre-computed, while F^k , F , W and $\mathbf{b}$ are the variables we want to solve. $\nu \in \mathbb{R}^+$, $\tau \in \mathbb{R}^+$, $\gamma \in \mathbb{R}^+$ and $\beta \in \mathbb{R}^+$ are the pre-defined parameters to balance the five terms. The first term in Eq. 5.21 utilizes the sub-manifold assumptions and imposes the smoothness constraint between elements in V_k . The second term is to ensure the label fitness so that the predicted label of every labeled image should comply with its original label. The third term enforces all the component-level label predictions in one image to be consistent with the image-level prediction label. We assign the i -th row of F^k the weight $\hat{m}_k^i$, which is the adapted image-specific weight of the k -th component in image i . The explanation is that we can regard the image-level label matrix F as the

Figure 5.3: Flowchart of the whole recognition process. Four steps of the process are indicated by red, blue, magenta and cyan rectangles respectively. These four steps are: 1) global GMM training (red rectangle); 2) MAP adaptation (blue rectangle); 3) PDC-SSDR (magenta rectangle); 4) label prediction (cyan rectangle).

weighted ensemble of $\{F^k\}_{k=1}^K$, and it is reasonable to assume that the components with

larger weights in image-specific GMMs contribute more to the image-level label matrix.

The fourth term is the residual between F and $X^T W + \mathbf{1}b^T$. Instead of using the hard constraint $F = X^T W + \mathbf{1}b^T$ as in LapRLS/L, we model the residual in our objective function to relax the hard constraint such that our method can cope with data from a non-linear manifold. The last term is a regularization term to control the complexity of the projection matrix.

5.3.3.3 Solution

Since Eq. 5.21 is an unconstrained quadratic programming problem with respect to F^k , F , W , and $\mathbf{b}$, we can obtain its closed form solution. Taking the derivatives of Eq. 5.21 with respect to $\mathbf{b}$ and setting it to zero, we have:

$$\begin{aligned}\frac{\partial \Omega}{\partial \mathbf{b}} &= 2\gamma \left(\mathbf{1}^T (X^T W + \mathbf{1} \mathbf{b}^T - F) \right) = 0 \\ \Rightarrow \mathbf{b} &= \frac{1}{l+u} (F^T \mathbf{1} - W^T X \mathbf{1}).\end{aligned}\quad (\text{Eq. 5.22})$$

Taking the derivatives of Eq. 5.21 with respect to F^k and setting it to zero, we have:

$$\begin{aligned}\frac{\partial \Omega}{\partial F^k} &= 2 \left(L^k F^k + \tau \Lambda^k (F^k - F) \right) = 0 \\ \Rightarrow F^k &= \tau (L^k + \tau \Lambda^k)^{-1} \Lambda^k F.\end{aligned}\quad (\text{Eq. 5.23})$$

Taking the derivatives of Eq. 5.21 with respect to W , setting it to zero and substituting $\mathbf{b}$ by Eq. 5.22, we have:

$$\begin{aligned}\frac{\partial \Omega}{\partial W} &= 2 \left(\gamma X (X^T W + \mathbf{1} \mathbf{b}^T - F) + \beta W \right) = 0 \\ \Rightarrow \gamma X (X^T W + \mathbf{1} \frac{1}{l+u} (F^T \mathbf{1} - W^T X \mathbf{1})^T - F) + \beta W &= 0 \\ \Rightarrow W &= (X A X^T + \frac{\beta}{\gamma} I)^{-1} X A F,\end{aligned}\quad (\text{Eq. 5.24})$$

where $A = (I - \frac{1}{l+u} \mathbf{1} \mathbf{1}^T)$. Taking the derivative of Eq. 5.21 with respect to F and substituting F^k , W and $\mathbf{b}$ by Eq. 5.23, Eq. 5.24, and Eq. 5.22 we have:

$$\begin{aligned}\frac{\partial \Omega}{\partial F} &= 2 \left(\nu U(F - Y) + \tau \sum_k \Lambda^k (F - F^k) - \gamma (X^T W + \mathbf{1} \mathbf{b}^T - F) \right) \\ &= 2 \left(\nu U(F - Y) + \tau \sum_k \Lambda^k (F - \tau (L^k + \tau \Lambda^k)^{-1} \Lambda^k F) \right. \\ &\quad - \gamma \left(X^T (X A X^T + \frac{\beta}{\gamma} I)^{-1} X A F \right. \\ &\quad \left. \left. + \mathbf{1} \frac{1}{l+u} \left(F^T \mathbf{1} - \left((X A X^T + \frac{\beta}{\gamma} I)^{-1} X A F \right)^T X \mathbf{1} \right)^T - F \right) \right).\end{aligned}\quad (\text{Eq. 5.25})$$

Setting Eq. 5.25 to zero, we can solve for the optimal F^* , from which F^{k*} , W^* and $\mathbf{b}^*$ can also be derived as:

$$\begin{aligned}
 F^* &= \nu M^{-1} UY, \\
 F^{k*} &= \tau(L^k + \tau\Lambda^k)^{-1} \Lambda^k F^* \\
 &= \tau\nu(L^k + \tau\Lambda^k)^{-1} \Lambda^k M^{-1} UY, \\
 W^* &= (XAX^T + \frac{\beta}{\gamma}I)^{-1} XAF^* \\
 &= \nu(XAX^T + \frac{\beta}{\gamma}I)^{-1} XAM^{-1} UY, \\
 \mathbf{b}^* &= \frac{1}{l+u} (F^{*T} \mathbf{1} - W^{*T} X \mathbf{1}). \tag{Eq. 5.26}
 \end{aligned}$$

where $M = \nu U + \tau \sum_k \Lambda^k - \tau^2 \sum_k \Lambda^k (L^k + \tau\Lambda^k)^{-1} \Lambda^k - \gamma AX^T (XAX^T + \frac{\beta}{\gamma}I)^{-1} XA + \gamma A$.

The outputs from PDC-SSDR are: 1) the prediction label matrix F^* , which also contains the prediction labels for the unlabeled training data; 2) the projection matrix W^* and bias $\mathbf{b}^*$, which can be used to obtain the lower-dimensional representation for any data point. In face recognition, for the unlabeled training data, we therefore have two choices: 1) directly use the following equation:

$$c_i = \underset{j}{\operatorname{argmax}} F_{ij}^*, i \in \{l+1, \dots, l+u\} \tag{Eq. 5.27}$$

for classification in face recognition, which was suggested in [40]; 2) use the Nearest Neighbor classifier for classification based on the lower-dimensional representations. Our experiments demonstrate that the results from Eq. 5.27 and NN are generally comparable on the unlabeled training data. For the unseen test data, we can only employ the Nearest Neighbor classifier for classification based on the lower-dimensional representation because we cannot obtain the prediction labels for the unseen test data during the training process. As human gait recognition is performed on videos, which are represented by a sequence of average silhouette images, we also use the simple fusion method

Figure 5.4: Sample images cropped from face databases CMU PIE (first row), FERET (second row) and AR (third and fourth row) respectively. The images in the third row and fourth row are from two sessions of the same person in the AR database.

proposed in Section 5.2.3 for recognition. For unlabeled training data, we can either use F^* or $X^T W^* + \mathbf{1b}^{*T}$ as the low-dimensional representations for recognition, while for the unseen test data, we can only represent them by $X^T W^* + \mathbf{1b}^{*T}$ as low-dimensional embeddings.

Figure 5.3 shows the whole process of recognition. The four steps of the process are indicated by red, blue, magenta and cyan rectangles respectively. These four steps are: 1) global GMM training (red rectangle); 2) MAP adaptation (blue rectangle); 3) PDC-SSDR (magenta rectangle); 4) label prediction (cyan rectangle).

The difference between PDC-SSDR and TR-FSDA, which is proposed in Chapter 4, is that PDC-SSDR is based on the LapRLS/L, which computes the projection matrix by linear regression, while TR-FSDA is based on SDA, which computes the projection matrix by eigenvalue decomposition. Although both PDC-SSDR and TR-FSDA relax the hard constraint of $F = X^T W + \mathbf{1b}^T$ (in TR-FSDA it is $F = X^T W$), PDC-SSDR utilizes the sub-manifold constraint, which is tailored for Patch Distribution Feature.

5.3.4 Experiments

To demonstrate the effectiveness of our PDC-SSDR algorithm, we test it on three benchmark face databases CMU PIE [2], FERET [1] and AR [21] as well as the USF HumanID gait database [3]. We compare our algorithm PDC-SSDR with the state-of-the-art semi-supervised learning algorithms SDA [6] and LapRLS/L [7], because these methods can handle unlabeled training samples as well as provide embeddings for unseen test samples by using the projection matrix W . It is desirable to also compare our previously proposed dimensionality reduction method TR-FSDA in Chapter 4 with PDC-SSDR as it is also a semi-supervised dimensionality reduction method. However, TR-FSDA is very slow on the Patch Distribution Feature (4K - 11K dimensions in our experiments, which are much higher than the gray-level features used in Chapter 4), as the process of TR-FSDA requires performing the time consuming eigenvalue decomposition iteratively, which results in much more processing time than SDA (SDA is much slower than LapRLS/L and PDC-SSDR when data dimension is high). In this section, we perform our experiments on the high-dimensional PDF directly (without the PCA preprocessing step), in which case, TR-FSDA is not practical. On the other hand, LapRLS/L and PDC-SSDR are much faster, and the experimental results of PDF on both face and human gait recognition can be easily achieved.

5.3.4.1 Face Recognition

Face images in all the databases are first aligned by manually fixing the positions of two eyes, and then cropped and resized with the gray-level values ranging from 0 to 255. Figure 5.4 shows some sample faces from the CMU PIE, FERET, and AR databases. These databases are introduced in detail in Section 2.1.1.

The CMU PIE (Pose, Illumination, and Expression) database [2] contains 41368 images of 68 subjects with 13 different poses, 43 different illumination conditions, and 4

different expressions. To make a fair comparison with [57] which also used CMU PIE, we adopt the same settings as [57] which only used a subset of images from the five near frontal poses (C05, C07, C09, C27, C29). There are 170 images per subject with all kinds of illuminations and expressions. Then we randomly select 1/5 of the samples out of 170 images, resulting in 34 images per subject. In total, there are 2312 images in this database. All images are resized to 20×20 pixels. For semi-supervised settings, we randomly select half of the images per subject as training samples, and use the rest of the images as test samples. Among the training images, two images are randomly selected as labeled samples and the rest are used as unlabeled samples.

A subset of the FERET [1] database which consists of 1400 images of 200 subjects is also used in our experiments. Every subject has seven images taken from seven different views, which is the same as the FERET database used in the experiments of Section 3.3 and Section 4.4. In this work, three images per subject are randomly selected as training data and the remaining four images are used as test data. Two out of the three training images are randomly selected as labeled data, and the remaining one image is used as unlabeled data. The images are resized to 36×36 pixels.

The AR [21] face database contains over 4,000 face images from 126 people (70 men and 56 women). A subset of 90 subjects with 26 images per subject are selected in our experiments and the images are resized to 31×21 pixels. All the images are frontal view faces with different facial expressions, illumination conditions, and occlusions (sunglasses and scarf). For each subject, the images are captured in two sessions separated by a two-week interval. In some challenging images, the glasses and scarf may occlude half of the face. We select 12 un-occluded images per subject from two sessions (the first 6 images of each session) as training data and use the remaining 14 images per subject as test data. Among the training images, two images are randomly selected as labeled samples and the remaining ten images are used as unlabeled samples.

We use two types of local patch features in our experiments for face recognition, the *normalized* feature and the 2-D Discrete Cosine Transform (DCT) feature. Similarly as Chapter 3, to extract the *normalized* features, local patches are first extracted from 3×3 windows centered at each pixel. Then the gray-level values inside each window are normalized to zero mean and unit standard deviation, resulting in H 9-dimensional *normalized* features, where H is the total number of pixels in the image. Illumination variation is reduced to some extent after the normalization process. For 2-D DCT feature, we then extract 2-D DCT coefficients from the *normalized* features, resulting in 9-dimensional DCT features. As suggested in [57], we further obtain 11-dimensional augmented local features by adding X-Y coordinates into *normalized* features and DCT features. Then, every image is represented by a set of 11-dimensional feature vectors from which the global GMM and image-specific GMMs are learned. As there are very few labeled images (two images in our experiments) per class for face recognition experiments, gallery expansion can only provide very limited virtual samples (one virtual sample per class), which leads to very little improvement to the recognition rate. Hence, we conduct our experiments on PDF without expansion directly.

We compare our PDC-SSDR algorithm with two state-of-the-art semi-supervised dimensionality reduction methods, SDA and LapRLS/L on three types of features (gray-level, *normalized* feature based PDF and DCT feature based PDF). We denote NN, SDA and LapRLS/L with gray-level features as Gray, Gray-SDA and Gray-LapRLS/L respectively. We refer to NN, SDA, LapRLS/L and PDC-SSDR using the *normalized* feature based PDF as PDF, PDF-SDA, PDF-LapRLS/L and PDC-SSDR respectively. The corresponding methods with DCT feature based PDF are denoted as DCT-PDF, DCT-PDF-SDA and DCT-PDF-LapRLS/L and DCT-PDC-SSDR respectively. For PDC-SSDR we also report the results by directly employing Eq. 5.27 for classifying the unlabeled training data (see Section 5.3.3.3). For each algorithm, we perform 10 rounds of random sample selection and report the average precision and standard deviation.

Figure 5.5: Performance variations (%) of Gray, Gray-SDA, Gray-LapRLS/L, PDF, PDF-SDA, PDF-LapRLS/L and PDC-SSDR on CMU PIE, FERET and AR face databases with respect to different values of K . The upper row shows the results for the unlabeled training data and the bottom row displays the results for the unseen test data. As Gray, Gray-SDA and Gray-LapRLS/L are not related to the PDF features, their recognition accuracies remain constant with respect to different values of K .

Several parameters need to be predefined for different methods. For SDA, we set the two parameters α and β both in the range of $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$, while for LapRLS/L we set γ_A and γ_I both in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4\}$ to ensure the algorithms reach their best results. While PDC-SSDR has four parameters, we observe that PDC-SSDR is relatively insensitive to the two parameters ν and τ . We therefore empirically fix $\nu = 10^9$ and $\tau = 10^{-2}$ and set the other two parameters γ and β both to values in the set $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$. For a fair comparison, we report the top-1 recognition accuracy from the best parameters for all the algorithms. It is still an open problem to automatically determine the optimal parameters in PDC-SSDR, which will be investigated in the future.

Another parameter in PDF, PDF-SDA, PDF-LapRLS/L and PDC-SSDR is the number of GMM components. We test PDF, PDF-SDA, PDF-LapRLS/L and PDC-SSDR on all the three face databases by setting $K = \{100, 200, 300, 400\}$ respectively, and plot the mean recognition accuracies and standard deviations of 10 rounds of experiments in Figure 5.5. We also plot Gray, Gray-SDA, and Gray-LapRLS/L in Figure 5.5 for the comparison of gray-level features and PDF. For a fair comparison, we report the recognition accuracies for PDF-SDA, PDF-LapRLS/L and PDC-SSDR on unlabeled training data using the NN classifier on lower-dimension representations in Figure 5.5. In practice, we observe that the results for PDC-SSDR on the unlabeled training data using Eq. 5.27 and NN are comparable. It can be observed from Figure 5.5 that the recognition accuracies on both unlabeled training data and unseen test data generally increase when K increases and then the results become stable at around $K = 400$. We also observe that PDC-SSDR consistently outperforms all the other algorithms on the three databases throughout all values of $K = \{100, 200, 300, 400\}$. In Table 5.2, we report the mean recognition accuracies and standard deviations of Gray, Gray-SDA, Gray-LapRLS/L, PDF, PDF-SDA, PDF-LapRLS/L and PDC-SSDR on CMU PIE, FERET and AR databases

with $K = 400$, in which the results for unlabeled training data and unseen test data are denoted by 'U' and 'T', respectively. For PDC-SSDR, the results from the two rows of the unlabeled training data are obtained by using Eq. 5.27 and NN respectively. The results in boldface are significantly better than the others judged by the t-test with a significance level of 0.05. From Table 5.2, we observe that PDC-SSDR significantly outperforms all the other methods on all three databases for both unlabeled training data and unseen test data. It is worth mentioning that the labeled and unlabeled training images on the AR database are all occlusion-free images while most of the unseen test images are with obvious occlusions, so the recognition accuracies on the unlabeled training data are much higher than those on unseen test data for all the algorithms. From Figure 5.5, it can also be observed that results of PDF (*resp.* PDF-SDA, and PDF-LapRLS/L) are better than Gray (*resp.* Gray-SDA, and Gray-LapRLS/L) for different values of K , which demonstrates the effectiveness of the PDF based patch distribution feature. Additionally, Gray-SDA and Gray-LapRLS/L outperform Gray, and PDF-SDA and PDF-LapRLS/L outperform PDF, which demonstrates that dimensionality reduction is useful in this application.

For DCT feature based PDF, we only perform experiments using $K = 400$ and compare the results with Table 5.2. The results are shown in Table 5.3. Results of DCT-PDF are similar to those of PDF using the *normalized* feature. However, the results of dimensionality reduction methods are all better. DCT-PDC-SSDR is still significantly better than all the other algorithms in almost all cases, which demonstrates the effectiveness of our PDC-SSDR using different types of features.

5.3.4.2 Human Gait Recognition

We also test our algorithm on the USF HumanID gait database collected by Sarkar et al. [3], which is currently the largest human gait database. The details of the database

Table 5.2: Top-one recognition rates (%) of Gray, Gray-SDA, Gray-LapRLS/L, PDF, PDF-SDA, PDF-LapRLS/L, and PDC-SSDR on the CMU PIE, FERET and AR databases using *normalized* features. For each data set, the results shown in boldface are significantly better than the others, judged by the t-test (with a significance level of 0.05).

Algorithm		CMU PIE [2]	FERET[1]	AR [21]
Gray	U	12.92 $\pm$ 0.72	20.85 $\pm$ 2.75	41.20 $\pm$ 2.31
	T	12.99 $\pm$ 1.06	21.40 $\pm$ 1.08	8.33 $\pm$ 0.74
Gray-SDA	U	38.33 $\pm$ 0.83	44.60 $\pm$ 2.73	73.94 $\pm$ 1.23
	T	39.12 $\pm$ 1.84	45.41 $\pm$ 1.66	23.90 $\pm$ 1.93
Gray-LapRLS/L	U	36.08 $\pm$ 1.03	48.75 $\pm$ 2.90	70.01 $\pm$ 1.59
	T	36.91 $\pm$ 1.54	49.56 $\pm$ 1.69	23.16 $\pm$ 2.57
PDF [57]	U	38.02 $\pm$ 1.30	50.75 $\pm$ 3.92	57.80 $\pm$ 1.24
	T	38.25 $\pm$ 1.56	52.20 $\pm$ 1.68	35.35 $\pm$ 1.22
PDF-SDA	U	49.88 $\pm$ 2.44	65.00 $\pm$ 5.26	81.27 $\pm$ 1.06
	T	51.99 $\pm$ 2.16	68.23 $\pm$ 1.86	46.55 $\pm$ 1.67
PDF-LapRLS/L	U	48.53 $\pm$ 2.41	59.05 $\pm$ 3.93	78.46 $\pm$ 1.61
	T	51.52 $\pm$ 1.82	68.69 $\pm$ 1.95	46.29 $\pm$ 1.71
PDC-SSDR	U	53.12 $\pm$ 2.34	70.80 $\pm$ 3.42	85.32 $\pm$ 0.82
		52.96 $\pm$ 2.44	70.65 $\pm$ 3.65	85.27 $\pm$ 0.68
	T	54.60 $\pm$ 1.98	71.53 $\pm$ 1.55	48.50 $\pm$ 1.67

Table 5.3: Top-one recognition rates (%) of DCT-PDF, DCT-PDF-SDA, DCT-PDF-LapRLS/L, and DCT-PDC-SSDR on the CMU PIE, FERET and AR databases. For each data set, the results shown in boldface are significantly better than the others, judged by the t-test (with a significance level of 0.05).

Algorithm		CMU PIE [2]	FERET[1]	AR [21]
DCT-PDF [57]	U	37.74 $\pm$ 1.59	50.50 $\pm$ 4.50	60.42 $\pm$ 1.79
	T	38.63 $\pm$ 1.51	52.48 $\pm$ 1.78	37.17 $\pm$ 1.42
DCT-PDF-SDA	U	54.03 $\pm$ 2.17	70.55 $\pm$ 3.40	85.21 $\pm$ 1.28
	T	55.08 $\pm$ 1.59	72.83 $\pm$ 1.48	52.05 $\pm$ 1.28
DCT-PDF-LapRLS/L	U	49.53 $\pm$ 2.07	61.25 $\pm$ 4.39	82.20 $\pm$ 1.46
	T	52.60 $\pm$ 1.50	71.48 $\pm$ 1.72	51.20 $\pm$ 1.44
DCT-PDC-SSDR	U	55.21 $\pm$ 1.89	73.95 $\pm$ 3.03	85.80 $\pm$ 1.07
		55.17 $\pm$ 1.80	73.75 $\pm$ 3.23	85.76 $\pm$ 1.06
	T	56.06 $\pm$ 1.69	74.06 $\pm$ 1.67	52.30 $\pm$ 1.40

is in Section 2.2.1. Following previous works [95, 101, 98, 100, 93, 125], we also use the average silhouette image of each cycle for gait recognition. The introduction of average silhouette images is in Section 2.2.3 and Section 3.3.2. Please refer to Figure 2.1 for the binary silhouette images and the corresponding average silhouette image.

Following our previous work in Section 5.2, we also use Gabor feature as the local patch feature for training the PDF and perform gallery expansion for experiments of human gait recognition. As demonstrated in Section 5.2.4, gallery expansion possesses huge advantage in human gait recognition, thus we also expand the gallery set in the experiments, which is denoted as PDF-EXP.

The size of each average silhouette image is 128×88 pixels. We extract 40-dimensional Gabor features from five scales and eight directions for each pixel and then down-sample each image to 64×44 . The same as Section 5.1.1, each image is represented by 2816 42-dimensional features after adding the X-Y coordinates to the 40-dimensional Gabor features. In our experiments, we set $K = 300$ for training the GMMs and we observe it is enough to achieve good results in all the experiments.

To compare semi-supervised learning algorithms on the USF HumanID database, we use the expanded gallery set as the labeled training data and select one probe set as the unlabeled training data. The remaining probe sets are employed as the unseen test data. Considering the performance may depend on the choice of the unlabeled training data, we repeat our experiments 12 times using each probe set as the unlabeled training data. So every probe set is used as unlabeled training data once and as test data 11 times. We record the recognition accuracies for each probe set as unlabeled training data and as the unseen test data separately. The reported accuracies for the probe sets used as test data are the average of 11 rounds of experiments. Again, the results for the unlabeled training data and the unseen test data are denoted by 'U' and 'T', respectively. Similar to face recognition, for PDC-SSDR the two rows for the unlabeled training data are the results

Table 5.4: Rank-1 recognition accuracy (%) comparison on the USF HumanID Gait database. Two sets of results are reported. The first 12 algorithms are supervised algorithms from previous works, while the last 7 methods are algorithms discussed in this section with 'U' and 'T' denoting the results for unlabeled data and test data respectively.

Probe Set		A	B	C	D	E	F	G	H	I	J	K	L	Average
Probe Size		122	54	54	121	60	121	60	120	60	120	33	33	–
Baseline [3]		73	78	48	32	22	17	17	61	57	36	3	3	40.96
HMM [119]		89	88	68	35	28	15	21	85	80	58	17	15	53.54
IMDE [111]		75	83	65	25	28	19	16	58	60	42	2	9	42.87
IMDE + LDA		88	86	72	29	33	23	32	54	62	52	8	13	48.64
LDA [95]		87	85	76	31	30	18	21	63	59	54	3	6	48.20
LDA + Sync [95]		83	94	61	50	48	22	33	48	52	34	18	12	48.04
LDA + Fusion [95]		91	94	81	51	57	25	29	62	60	57	9	12	55.83
DNGR [92]		85	89	72	57	66	46	41	83	79	52	15	24	62.81
MMFA [100]		89	94	80	44	47	25	33	85	83	60	27	21	59.9
GTDA [101]		91	93	86	32	47	21	32	95	90	68	16	19	60.58
Image-to-Class [125]		93	89	81	54	52	32	34	81	78	62	12	9	61.19
PDF-EXP-RTRDA		95	91	85	59	62	36	36	84	76	73	3	21	65.41
Gray		83	85	67	24	24	9	16	51	47	42	0	6	40.61
Gray-SDA	U	87	85	74	36	34	15	26	81	59	59	6	9	51.99
	T	87	85	76	35	34	16	26	60	56	58	4	10	49.12
Gray-LapRLS/L	U	89	87	78	32	34	18	26	70	53	60	6	12	50.93
	T	89	87	77	30	37	18	29	58	52	58	5	12	49.15
PDF-EXP [57]		93	87	81	48	55	30	22	81	76	63	6	18	59.61
PDF-EXP-SDA	U	95	87	83	53	53	31	33	83	72	69	3	6	61.54
	T	95	88	83	53	54	30	34	81	72	69	3	6	61.34
PDF-EXP-LapRLS/L	U	94	91	85	57	62	23	40	85	81	64	15	15	63.15
	T	93	92	84	55	59	33	36	88	79	67	14	14	64.16
EXP-PDC-SSDR	U	98	93	89	76	79	45	52	96	91	82	12	24	75.46
		97	94	89	76	78	47	52	95	90	79	18	21	75.12
	T	95	91	87	67	69	40	43	88	85	72	9	18	68.96

Table 5.5: Rank-5 recognition accuracy (%) comparison on the USF HumanID Gait database. Two sets of results are reported. The first 12 algorithms are supervised algorithms from previous works, while the last 7 methods are algorithms discussed in this section with 'U' and 'T' denoting the results for unlabeled data and test data respectively.

Probe Set		A	B	C	D	E	F	G	H	I	J	K	L	Average
Probe Size		122	54	54	121	60	121	60	120	60	120	33	33	–
Baseline [3]		88	93	78	66	55	42	38	85	78	62	12	15	64.54
HMM [119]		–	–	–	–	–	–	–	–	–	–	–	–	–
IMDE [111]		91	93	83	52	59	41	38	86	76	76	12	15	65.31
IMDE + LDA		95	95	90	52	63	42	47	86	86	78	21	19	68.60
LDA [95]		92	93	89	58	60	36	43	90	81	79	12	12	67.37
LDA + Sync [95]		92	96	91	68	69	50	55	80	78	69	39	30	70.85
LDA + Fusion [95]		94	96	93	85	79	52	57	89	86	77	24	21	76.18
DNGR [92]		96	94	89	85	81	68	69	96	95	79	46	39	82.05
MMFA [100]		98	98	94	76	76	57	60	95	93	84	48	39	79.9
GTDA [101]		98	99	97	68	68	50	56	95	99	84	40	40	77.58
Image-to-Class [125]		97	98	93	81	74	59	55	94	95	83	30	33	79.17
PDF-EXP-RTRDA		98	94	96	79	79	64	60	95	93	86	18	30	80.11
Gray		93	93	91	57	52	31	34	79	79	69	12	15	63.13
Gray-SDA	U	95	93	93	63	59	42	47	89	83	80	21	18	70.19
	T	95	93	92	63	58	42	47	86	82	80	21	20	69.71
Gray-LapRLS/L	U	94	94	93	58	62	44	48	91	86	81	24	18	70.66
	T	94	94	91	60	61	43	48	84	83	80	24	19	69.46
PDF-EXP [57]		98	94	91	75	72	49	48	95	98	82	24	36	76.46
PDF-EXP-SDA	U	98	96	93	84	83	58	55	92	91	86	21	27	79.36
	T	98	96	93	82	82	59	56	92	91	85	20	27	79.08
PDF-EXP-LapRLS/L	U	97	93	94	85	86	59	64	97	98	84	42	33	81.87
	T	97	93	94	83	86	62	62	95	97	85	45	31	81.72
EXP-PDC-SSDR	U	98	98	94	92	91	75	81	99	98	95	45	42	88.60
		99	98	94	91	91	73	74	99	98	91	48	39	87.41
	T	98	94	94	89	91	70	71	97	98	88	37	32	84.99

obtained by using F^* and $X^T W^* + \mathbf{1b}^{*T}$ as low-dimensional representation, respectively. We also use simple fusion to do classification. After obtaining the recognition accuracy of each probe set, we compute the *average* recognition accuracy of all probe sets according to [100], which is the ratio of correctly recognized people to the total number of people in all the 12 probe sets.

We report results of Gray, Gray-SDA, Gray-LapRLS/L, PDF-EXP, PDF-EXP-SDA, PDF-EXP-LapRLS/L and EXP-PDC-SSDR on all the probe sets and compare them with existing works. As Gray and PDF are supervised learning algorithms, we only report the results of each probe set by using it as the test data. Table 5.4 reports the Rank-1 recognition accuracy and Table 5.5 reports the Rank-5 recognition accuracy. Again, PDF-EXP, PDF-EXP-SDA and PDF-EXP-LapRLS/L outperform Gray, Gray-SDA and Gray-LapRLS/L respectively, demonstrating the effectiveness of PDF features and gallery expansion. Our best results for the unlabeled training data are 75.46% (Rank-1) and 88.60% (Rank-5), which are much higher than the second best results 63.15% (Rank-1) and 81.87% (Rank-5) among all the semi-supervised algorithms. The best results for the unseen test data are 68.96% (Rank-1) and 84.99% (Rank-5), which are also higher than the second best results 64.16% (Rank-1) and 81.72% (Rank-5) among all the semi-supervised algorithms. It clearly demonstrates the effectiveness of our algorithm by using the sub-manifold smoothness assumption and modeling the regression residual.

We also report recognition results from previous supervised algorithms, which are similar to Table 5.1. Note that the semi-supervised algorithms use more data for training than the supervised algorithms in the literature. However, only one probe set is used as the unlabeled training data. Some supervised algorithms outperform Gray-SDA and Gray-LapRLS/L by utilizing better features or gait dynamics, demonstrating that such a limited amount of unlabeled training data is not enough for SDA and LapRLS/L to beat all the supervised algorithms. The results show that PDF features with gallery

expansion improve the performance from the Gray 40.61% to PDF-EXP 59.61% (Rank-1), and from 63.13% to 76.46% (Rank-5). PDF-EXP-SDA and PDF-EXP-LapRLS/L are comparable to the best supervised algorithm in terms of the average Rank-1 and Rank-5 recognition accuracies, demonstrating the effectiveness of the semi-supervised learning methods in combination with the PDF. Our own work PDF-EXP-RTRDA proposed in Section 5.2, which uses the same feature as EXP-PDC-SSDR, achieves the best Rank-1 recognition rate while DNCR [92] achieves the best Rank-5 recognition rate among all the existing supervised algorithms. EXP-PDC-SSDR outperforms both PDF-EXP-RTRDA and DNCR in terms of Rank-1 and Rank-5 recognition accuracies, demonstrating the effectiveness of our proposed algorithm.

5.3.5 Conclusion

In this section we propose a new semi-supervised dimensionality reduction algorithm called Patch Distribution Compatible Semi-Supervised Dimensionality Reduction (PDC-SSDR) for face and human gait recognition. First, we represent every image by its Patch Distribution Feature which can be expressed by an image-specific Gaussian Mixture Model (GMM) adapted from the Universal Background Model. Second, instead of assuming that the high-dimensional Patch Distribution Features reside on one manifold, we assume that there are a set of sub-manifolds formed by individual GMM components which represent the appearance features extracted from specific facial areas. Third, we relax the over-strict constraint $F = X^T W + \mathbf{1b}^T$ in Linear Laplacian Regularized Least Squares (LapRLS/L) by adding a residual term $\|X^T W + \mathbf{1b}^T - F\|^2$ in our objective function. Utilizing the sub-manifold smoothness assumption and modeling the residual term in our objective function, the component-level label matrix F^k , the image-level label matrix F , the projection matrix W , and the bias term $\mathbf{b}$ can be solved simultaneously. Experiments on three face databases and one human gait database show that the

proposed algorithm performs better than the state-of-the-art semi-supervised dimensionality reduction methods SDA and LapRLS/L. In the future, we will investigate how to automatically determine the optimal parameters in PDC-SSDR.

Chapter 6

Conclusion

This thesis addresses the issues in feature extraction/processing step and classification step of face recognition and human gait recognition. For feature extraction/processing part, we propose several new dimensionality reduction and graph embedding methods based on either gray-level features or the Patch Distribution Features to map the high-dimensional data to low-dimensional spaces in both supervised and semi-supervised settings. For the classification step, we also introduce the new enhanced Image-to-Class distance for measuring the distance between probe images and gallery images.

6.1 Summary of Contributions

- We propose a new distance measure for face recognition and human gait recognition. Each probe image (a face image or an average human silhouette image) is represented as a set of local features uniformly sampled over a grid with fixed spacing, and each gallery image is represented as a set of local features sampled at each pixel. We formulate an integer programming problem to compute the distance (referred to as Image-to-Class distance) from one probe image to all the the gallery images belonging to a certain class, in which any feature of the probe image can be matched to only one feature from one of the gallery images. Considering computational efficiency as well as the fact that face images or average human silhouette images are

roughly aligned in the preprocessing step, we also enforce a spatial neighborhood constraint by only allowing neighboring features that are within a given spatial distance to be considered for feature matching. The integer programming problem is further treated as a classical minimum-weight bipartite graph matching problem, which can be efficiently solved with the Kuhn-Munkres algorithm. We performed comprehensive experiments on three benchmark face databases CMU PIE, FERET and FRGC as well as the USF HumanID gait database. The experiments clearly demonstrate the effectiveness of our Image-to-Class distance.

- A new semi-supervised dimensionality reduction method called Trace Ratio based Flexible Semi-supervised Discriminant Analysis (TR-FSDA) is proposed. In many dimensionality reduction algorithms (*e.g.*, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Semi-supervised Discriminant Analysis (SDA)), the low-dimensional representation F is constrained to be in the linear subspace spanned by the data matrix X , namely, $F = X^T W$, in which the projection matrix W can be used to map the unseen test data. In order to better cope with the data points sampled from nonlinear manifold, in this work we first relax this hard constraint $F = X^T W$ by introducing a flexible regularizer $\|F - X^T W\|^2$ into the objective function of SDA. Based on the recent study that trace ratio based LDA generally outperforms ratio trace based LDA, we further formulate our objective function in a trace ratio form. As there is no closed form solution to our new objective function, we propose an iterative algorithm to solve it and we prove that our iterative algorithm converges to the optimum based on the fast Newton-Raphson method. Extensive experiments on three face databases demonstrate the effectiveness of TR-FSDA.
- Then we propose a new dimensionality reduction algorithm in combination with the Gaussian Mixture Model (GMM) based Patch Distribution Feature for human gait

recognition. Instead of representing each average silhouette image as its gray-level feature, we first extract local patch features at every pixel of the average silhouette image and train a GMM to describe the distribution of the patches in each image. A Universal Background Model (UBM) is first trained with local patch features from all gallery images, then every gallery or probe image is represented by the distribution parameters (referred to as Patch Distribution Features (PDF)) of the image-specific GMM adapted from the UBM. To cope with the high dimension of the PDF feature, the Regularized Trace Ratio Discriminant Analysis (RTRDA) is developed to find the most discriminative subspaces for gait recognition. Experiments on USF humanID database show that RTRDA significantly outperforms the existing algorithms and achieves the best recognition results among all the previous works on USF humanID database in terms of average Rank-1 recognition rate.

- A new semi-supervised learning algorithm, referred to as Patch Distribution Compatible Semi-Supervised Dimensionality Reduction (PDC-SSDR), is also proposed for face and human gait recognition. Each image (a face image or an average human silhouette image) is first represented as a set of local Patch features and it is further characterized as the corresponding Patch Distribution Feature. Assuming that the individual components of the image-specific GMMs from all the training images reside on a sub-manifold, we assign a component-level prediction label matrix F^k to individual components and introduce a new regularizer based on a set of local sub-manifold smoothness assumptions in our objective function. We also constrain the component-level prediction label matrix F^k to be consistent with the image-level prediction label matrix F , as well as enforce F to be close to the given labels for the labeled samples. We further use a linear regression function $X^T W + \mathbf{1} \mathbf{b}^T$ to provide embeddings for the unseen test data, and additionally integrate the regression residual $\|X^T W + \mathbf{1} \mathbf{b}^T - F\|^2$ in our objective function to measure the mismatch

between $X^T W + \mathbf{1b}^T$ and F . By modeling the regression residual, we show that we can relax the over-strict hard constraint $F = X^T W + \mathbf{1b}^T$ in the multidimensional version of Linear Laplacian Regularized Least Squares (LapRLS/L) such that our method can better cope with the data sampled from a non-linear manifold. Finally, the optimal solutions of the component-level label matrix F^k , the image-level label matrix F , the projection matrix W , and the bias term $\mathbf{b}$ can be simultaneously obtained. Comprehensive experiments on three benchmark face databases CMU PIE, FERET and AR as well as the USF HumanID gait database clearly demonstrate the effectiveness of our algorithm over other state-of-the-art semi-supervised dimensionality reduction methods.

6.2 Future Research Directions

There exist some issues in feature extraction/processing and classification of face recognition and human gait recognition which deserve future research attempts.

- In Image-to-Class distance, l_2 -norm is used to measure the distance between local patch features and then the distance computation problem is formulated the as an integer programming problem. Recently, sparse representation is very popular in face recognition, in which each probe images is linearly reconstructed by as few gallery images as possible. The problem is formulated in l_1 -norm. Then the reconstruction error is used as the distance between the probe image and the gallery images for classification. In the previous work of face recognition using l_1 -norm [127], each image is represented by holistic features composed by all the pixel values. In the future, the l_1 -norm can also be applied as the distance measure between local patch features to compute the Image-to-Class or Image-to-Image distance. As the integer programming only allows one probe local patch feature to be matched to

one gallery local patch feature, sparse representation can relax this hard constraint by allowing a few gallery local patch features to linearly reconstruct the probe local patch feature. Then the distance between two local patch features is represented by the reconstruction error of the sparse representation. When there are a set of gallery images with different poses of the same person, linear reconstruction from several images with sparse representation may provide better matching than the one-one matching imposed by integer programming. The Image-to-Class distance can also be integrated with other dimensionality reduction methods.

- As Patch Distribution Feature (PDF) is demonstrated very effective in face and human gait recognition, how to improve the learning process for building the PDF is also worth of further research. Work can be done to automatically determine the optimal number of components in the Gaussian Mixture Model (GMM) to describe the set of local features. While dimensionality reduction algorithms are always proposed for holistic features, they may also be applied to local patch features. Many dimensionality reduction methods adopt the data distribution information to find the best projection directions for data mapping, hence in the process of building the global GMM and the image-specific GMMs, we can employ dimensionality reduction methods to project the local patch features into lower-dimensional space where they can be better clustered into different GMM components, resulting in better PDF for describing the patch distributions of each image.
- There have been numerous endeavors which try to learn good descriptors to characterize the set of local features and design good measures for two sets of data. PDF is based on GMM, which assumes the data distribution is Gaussian, can be seen as one of them. New PDF can be designed to better capture the distribution of the local feature set, utilizing other distribution assumptions or non-parametric distributions.

- As there are several parameters in dimensionality reduction methods, how to automatically determine the values of these parameters is also a issue which deserves further consideration. Additionally, in Chapter 4, we relax the hard constraint $F = X^T W$ and add the residue term $\|F - X^T W\|^2$ in the objective function of Semi-supervised Discriminant Analysis (SDA) to propose our Trace Ratio based Flexible Discriminant Analysis (TR-FSDA). The residue term may also be applicable to other discriminant analysis methods, under different criteria for mapping data and separating data.

References

- [1] P. Philips, S. Rizvi H. Moon, and P. Rauss, “The feret evaluation methodology for face recognition algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, Oct 2000.
- [2] T. Sim and S. Baker, “The cmu pose, illumination and expression database,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1617, Dec 2003.
- [3] S. Sarkar, P. Phillips, Z. Liu, I. Vega, P. Grother, and K. 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, Feb 2005.
- [4] P. Belhumeur, J. Hespanha, and D. 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, Jul 1997.
- [5] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, “Trace ratio vs. ratio trace for dimensionality reduction,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA, 2007, pp. 1 – 8.
- [6] D. Cai, X. He, and J. Han, “Semi-supervised discriminant analysis,” in *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1–7.
- [7] V. Sindhwani, P. Niyogi, M. Belkin, and S. Keerthi, “Linear manifold regularization for large scale semi-supervised learning,” in *International Conference on Machine*

REFERENCES

- Learning (ICML) Workshop on Learning with Partially Classified Training Data*, Bonn, Germany, 2005.
- [8] W. Liu, D. Tao, and J. Liu, “Transductive component analysis,” in *IEEE International Conference on Data Mining*, Pisa, Italy, 2008, pp. 433–442.
- [9] F. Nie, D. Xu, I. Tsang, and C. Zhang, “Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension,” *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1921 – 1932, Jul 2010.
- [10] Wikipedia, <http://en.wikipedia.org/wiki/Biometrics>.
- [11] T. Cootes, G. Edwards, and C. Taylor, “Comparing active shape models with active appearance models,” in *British Machine Vision Conference*, Nottingham, UK, 1999, pp. 173–182.
- [12] P. Viola and M. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.
- [13] D. Xu, S. Lin, S. Yan, and X. Tang, “Rank-one projections with adaptive margins for face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, 2006, pp. 175 – 181.
- [14] D. Xu, S. Yan, and J. Luo, “Face recognition using spatially constrained earth mover’s distance,” *IEEE Transactions on Image Processing*, vol. 17, no. 11, pp. 2256–2260, Nov 2008.
- [15] G. Johansson, “Visual motion perception,” *Scientific American*, vol. 232, no. 6, pp. 76–88, 1975.
- [16] S. V. Stevenage, M. S. Nixon, and K. Vince, “Visual analysis of gait as a cue to identity,” *Applied Cognitive Psychology*, vol. 13, no. 6, pp. 513–526, 1999.
- [17] M. W. Whittle, “Clinical gait analysis: A review,” *Human Movement Science*, vol. 15, no. 3, pp. 369–387, Jun 1996.

REFERENCES

- [18] R. O. Duda, P. E. Hart, and D. H. Stork, *Pattern Classification (Second Edition)*, Wiley Interscience, Hoboken, NJ, USA, 2000.
- [19] D. B. Graham and N. M. Allinson, “Characterizing virtual eigensignatures for general purpose face recognition,” *Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and Systems Sciences*, vol. 163, pp. 446–456, 1998.
- [20] P. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, M. Jaesik, and W. Worek, “Overview of the face recognition grand challenge,” in *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 947–954.
- [21] A. M. Martinez and R. Benavente, “The ar face database,” in *CVC Technical Report 24*, 1998.
- [22] M. Turk and A. Pentland, “Face recognition using eigenfaces,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Maui, Hawaii, USA, 1991, pp. 586–591.
- [23] Y. Guo, S. Li, J. Yang, T. Shu, and L. Wu, “A generalized foley-sammon transform based on generalized fisher discriminant criterion and its application to face recognition,” *Pattern Recognition Letters*, vol. 24, no. 1-3, pp. 147–158, 2003.
- [24] Y. Jia, F. Nie, and C. Zhang, “Trace ratio problem revisited,” *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, Apr 2009.
- [25] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” in *Advances in Neural Information Processing Systems 14*, Vancouver, British Columbia, Canada, 2001, pp. 585–591.
- [26] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec 2000.
- [27] J. B. Tenenbaum, V. Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec 2000.

REFERENCES

- [28] X. He, D. Cai, S. Yan, and H. J. Zhang, “Neighborhood preserving embedding,” in *IEEE International Conference on Computer Vision*, Kyoto, Japan, 2005, pp. 1208–1213.
- [29] X. He and P. Niyogi, “Locality perserving projections,” in *Advances in Neural Information Processing Systems*, Cambridge, MA, USA, 2004.
- [30] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, “Face recognition using laplacian-faces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, Mar 2005.
- [31] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, “Graph embedding and extension: A general framework for dimensionality reduction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, Jan 2007.
- [32] 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, Mar 2008.
- [33] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, “Learning with l_1 -graph for image analysis,” *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 858–866, Apr 2010.
- [34] Y. Fu, S. Yan, and T. S. Huang, “Correlation metric for generalized feature extraction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2229–2235, Dec 2008.
- [35] Y. Fu, Z. Li, J. Yuan, Y. Wu, and T. S. Huang, “Locality versus globality: Query-driven localized linear models for facial image computing,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 12, pp. 1741–1752, Dec 2008.
- [36] Y. Fu and T. S. Huang, “Manifolds and subspace learning for pattern recognition,” *Pattern Recognition and Machine Vision - In Honor and Memory of Professor King-Sun Fu, Patrick Shen-pei Wang (Ed.)*, pp. 215–229, 2010.

REFERENCES

- [37] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *COLT’ 98: Proceedings of the eleventh annual conference on Computational learning theory*, Madison, Wisconsin, USA, 1998, pp. 92–100.
- [38] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, Hoboken, NJ, USA, 1998.
- [39] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples,” *The Journal of Machine Learning Research*, vol. 7, no. 12, pp. 2399–2434, Dec 2006.
- [40] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, “Learning with local and global consistency,” Vancouver, B.C., Canada, 2004.
- [41] X. Zhu, Z. Ghahramani, and J. Lafferty, “Semi-supervised learning using gaussian fields and harmonic functions,” in *International Conference on Machine Learning*, Washington, DC, USA, 2003, pp. 912–919.
- [42] V. Sindhwani and P. Niyogi, “Beyond the point cloud: from transductive to semi-supervised learning,” in *International Conference on Machine Learning*, Bonn, Germany, 2005, pp. 824–831.
- [43] 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.
- [44] J. Ye, R. Janardan, and Q. Li, “Two-dimensional linear discriminant analysis,” in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, B.C. Canada, 2004, pp. 1569–1576.
- [45] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. J. Zhang, “Discriminant analysis with tensor representation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 526–532.

REFERENCES

- [46] D. Xu, S. Lin, S. Yan, and X. Tang, “Rank-one projections with adaptive margins for face recognition,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 5, pp. 1226 –1236, Oct 2007.
- [47] T. Kanade, *Computer Recognition of Human Faces*, Birkhauser Verlag, Stuttgart, Germany, 1977.
- [48] M. D. Kelly, *Visual identification of people by computer*, Ph.D. thesis, 1971.
- [49] F. Samaria and S. Young, “Hmm-based architecture for face identification,” *Image and Vision Computing*, vol. 12, no. 8, pp. 537 – 543, 1994.
- [50] M. D. Kelly, *Visual identification of people by computer*, Ph.D. thesis, 1971.
- [51] J. Buhmann, M. Lades, and C. von der Malsburg, “Size and distortion invariant object recognition by hierarchical graph matching,” in *International Joint Conference on Neural Networks (IJCNN)*, Washington, D.C., USA, 1990, pp. 411 –416.
- [52] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. V. D. Malsburg, R. P. Wrtz, and W. Konen, “Distortion invariant object recognition in the dynamic link architecture,” *IEEE Transactions on Computers*, vol. 42, no. 3, pp. 300–311, Mar 1993.
- [53] L. Wiskott, J. M. Fellous, N. Kuiger, and C. von der Malsburg, “Face recognition by elastic bunch graph matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, Jul 1997.
- [54] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. Malsburg, “The bochum/usc face recognition system and how it fared in the feret phase iii test,” 1998.
- [55] S. Lucey and T. Chen, “A gmm parts based face representation for improved verification through relevance adaptation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC., USA, 2004, vol. 2, pp. 855–861.
- [56] S. Lucey and T. Chen, “Using overlapping distributions to deal with face pose mismatch,” in *British Machine Vision Conference*, Oxford Brookes, UK, 2005.

- [57] M. Liu, S. Yan, Y. Fu, and T. S. Huang, “Flexible x-y patches for face recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, 2008, pp. 2113 – 2116.
- [58] J. Zhang, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: a comprehensive study,” *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213 – 238, Jun 2007.
- [59] J. Huang, V. Blanz, and B. Heisele, “Face recognition using component-based svm classification and morphable models,” in *International Workshop on Pattern Recognition with Support Vector Machines*, Niagara Falls, Canada, 2002, pp. 334–341.
- [60] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, “On the use of sift features for face authentication,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, New York, NY, USA, 2006, p. 35.
- [61] Z. Zhou, A. Ganesh, J. Wright, S. F. Tsai, and Y. Ma, “Nearest-subspace patch matching for face recognition under varying pose and illumination,” in *IEEE International Conference on Automatic Face & Gesture Recognition*, Amsterdam, The Netherlands, 2008, pp. 1–8.
- [62] H. Wang, S. Yan, T. Huang, J. Liu, and X. Tang, “Misalignment-robust face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1 – 6.
- [63] A. B. Ashraf, S. Lucey, and T. Chen, “Learning patch correspondences for improved viewpoint invariant face recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.
- [64] A. Li, S. Shan, X. Chen, and W. Gao, “Maximizing intra-individual correlations for face recognition across pose differences,” Miami, FL, USA, 2009, pp. 605–611.
- [65] J. Wright and G. Hua, “Implicit elastic matching with random projections for pose-variant face recognition,” Miami, FL, USA, 2009, pp. 1502–1509.

- [66] R. Kumar, A. Banerjee, and B. C. Vemuri, “Volterrafaces: Discriminant analysis using volterra kernels,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 150 –155.
- [67] Z. Cao, Q. Yin, J. Sun, and X. Tang, “Face recognition with learning-based descriptor,” in *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 2010, pp. 2707 – 2714.
- [68] Z. Wu, Q. Ke, J. Sun, and H. Y. Shum, “Scalable face image retrieval with identity-based quantization and multi-reference re-ranking,” in *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 2010, pp. 3469 – 3476.
- [69] J. Cutting and L. Kozlowski, “Recognizing friends by their walk: Gait perception without familiarity cues,” *Bulletin of the Psychonomic Society*, vol. 9, pp. 353–356, 1977.
- [70] L. Kozlowski and J. Cutting, “Recognizing the sex of a walker from a dynamic point-light display,” *Perception and Psychophysics*, vol. 21, pp. 575–580, 1977.
- [71] J. Cutting and D. Proffitt, “Gait perception as an example of how we may perceive events,” *Intersensory Perception and Sensory Integration*, pp. 249–273, 1981.
- [72] A. F. Bobick and A. Y. Johnson, “Gait recognition using static activity-specific parameters,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, USA, 2001, pp. 423–430.
- [73] J. Little and J. E. Boyd, “Recognizing people by their gait: The shape of motion,” *Journal of Computer Vision Research*, vol. 1, no. 2, pp. 2–32, 1998.
- [74] M. S. Nixon, J. N. Carter, J. M. Nash, P. S. Huang, D. Cunado, and S. V. Stevenage, “Automatic gait recognition,” in *IEE Colloquium on Motion Analysis and Tracking*, London , UK, 1999, pp. 1–6.
- [75] H. Murase and R. Sakai, “Moving object recognition in eigenspace representation: gait analysis and lip reading,” *Pattern Recognition Letters*, vol. 17, no. 2, pp. 155–162, 1996.

REFERENCES

- [76] S. A. Niyogi and E. H. Adelson, “Analyzing and recognizing walking figures in xyt,” in *IEEE Conference on Computer Vision and Patter Recognition*, Sarasota, FL, USA, 1994, pp. 469–474.
- [77] S. L. Darrell, G. Shakhnarovich, L. Lee, and T. Darrell, “Integrated face and gait recognition from multiple views,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, USA, 2001, pp. 439–446.
- [78] R. Tanawongsuwan and A. F. Bobick, “Gait recognition from time-normalized joint-angle trajectories in the walking plane.,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Kauai, HI, USA, 2001, pp. 726–731.
- [79] R. Gross and J. Shi, “The cmu motion of body (mobo) database,” Tech. Rep., Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001.
- [80] A. Y. Johnson and A. F. Bobick, “A multi-view method for gait recognition using static body parameters,” in *International Conference on Audio- and Video-Based Biometric Person Authentication*, Halmstad, Sweden, 2001, pp. 301–311.
- [81] A. Kale, N. Cuntoor, and R. Chellappa, “A multi-view method for gait recognition using static body parameters,” in *International Conference on Acoustics Speech, and Signal Processing*, Orlando, FL, USA, 2002.
- [82] N. Cuntoor, A. Kale, and R. Chellappa, “Combining multiple evidences for gait recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, China, 2003, pp. 6–10.
- [83] J. B. Hayfron-acquah, M. S. Nixon, and J. N. Carter, “Automatic gait recognition by symmetry analysis,” *Pattern Recognition Letters*, vol. 24, pp. 2175–2183, 2003.
- [84] Chinese Academy of Sciences (CASIA) Gait Database, [http://www.cbsr.ia.ac.cn/english/Gait Databases.asp](http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp).
- [85] S. Carlsson, “Recognizing walking people,” in *European Conference on Computer Vision*, Dublin, Ireland, 2000, pp. 472–486.

REFERENCES

- [86] L. Lee, “Gait analysis for recognition and classification,” in *IEEE Conference on Face and Gesture Recognition*, Washington, D.C., USA, 2002, pp. 155–161.
- [87] R. Tanawongsuwan and A. Bobick, “Modelling the effects of walking speed on appearance-based gait recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA, 2004, pp. 783–790.
- [88] R. Urtasun and P. Fua, “3d tracking for gait characterization and recognition,” in *IEEE Automatic Face and Gesture Recognition*, Seoul, Korea, 2004, pp. 17–22.
- [89] S. L. Dockstader, M. J. Berg, and A. M. Tekalp, “Stochastic kinematic modeling and feature extraction for gait analysis,” *Image Processing, IEEE Transactions on*, vol. 12, no. 8, pp. 962 – 976, Aug 2003.
- [90] C. B. Abdelkader, “Stride and cadence as a biometric in automatic person identification and verification,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, Washington, D.C., USA, 2002, p. 372.
- [91] L. Wang, H. Ning, T. Tan, and W. Hu, “Fusion of static and dynamic body biometrics for gait recognition,” *IEEE Transactions on Circuits and Systems for Video Technology Special Issue on Image- and Video-Based Biometrics*, vol. 14, no. 2, pp. 149–158, Mar 2004.
- [92] Z. Liu and S. Sarkar, “Improved gait recognition by gait dynamics normalization,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 863–876, Jun 2006.
- [93] Z. Liu and S. Sarkar, “Simplest representation yet for gait recognition: Averaged silhouette,” in *IEEE International Conference on Pattern Recognition*, Cambridge, England, UK, 2004, pp. 211 – 214.
- [94] J. Han and B. Bhanu, “Statistical feature fusion for gait-based human recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA, 2004, pp. 842–847.

REFERENCES

- [95] J. Han and B. Bhanu, “Individual recognition using gait energy image,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–332, Feb 2006.
- [96] 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, Dec 2003.
- [97] L. Wang, T. Tan, W. Hu, and H. Ning, “Automatic gait recognition based on statistical shape analysis,” *IEEE Transactions on Image Processing*, vol. 12, no. 9, pp. 1120 – 1131, Sep 2003.
- [98] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, and H. Zhang, “Human gait recognition with matrix representation,” *IEEE Transactions on Circuits Systems for Video Technology*, vol. 16, no. 7, pp. 896–903, Jul 2006.
- [99] D. Xu, S. Yan, L. Zhang, H. J. Zhang, Z. Liu, and H. Shum, “Concurrent subspaces analysis,” in *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 203–208.
- [100] D. Xu, S. Yan, D. Tao, S. Lin, and H. Zhang, “Marginal fisher analysis and its variants for human gait recognition and content based image retrieval,” *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2811–2821, Nov 2007.
- [101] D. Tao, X. Li, X. Wu, and S. J. Maybank, “General tensor discriminant analysis and gabor features for gait recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700–1715, Oct 2007.
- [102] X. Li, S. Lin, S. Yan, and D. Xu, “Discriminant locally linear embedding with high-order tensor data,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 2, pp. 342 –352, Apr 2008.
- [103] C. Chen, J. Zhang, and R. Fleischer, “Distance approximating dimension reduction of riemannian manifolds,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 1, pp. 208 –217, Feb 2010.

REFERENCES

- [104] 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, Mar 2008.
- [105] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, “A study on gait-based gender classification,” *IEEE Transactions on Image Processing*, vol. 18, no. 8, pp. 1905–1910, Aug 2009.
- [106] R. Chellappa, A. K. Roy-Chowdhury, and A. Kale, “Human identification using gait and face,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA, 2007, pp. 1–2.
- [107] C. Shan, S. Gong, and P. W. McOwan, “Fusing gait and face cues for human gender recognition,” *Neurocomputing*, vol. 71, no. 10-12, pp. 1931–1938, 2008.
- [108] S. Z. Li and J. Lu, “Face recognition using the nearest feature line method,” *IEEE Transactions on Neural Networks*, vol. 10, no. 2, pp. 439–443, Mar 1999.
- [109] 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, Dec 2002.
- [110] R. Wang, S. Shan, X. Chen, and W. Gao, “Manifold-manifold distance with application to face recognition based on image set,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, Jun 2008, pp. 1–8.
- [111] L. Wang, Y. Zhang, and J. Feng, “On the euclidean distance of images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1334–1339, Aug 2005.
- [112] O. Boiman, E. Shechtman, and M. Irani, “In defense of nearest-neighbor based image classification,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

REFERENCES

- [113] L. Chen, L. Zhang, L. Zhu, M. Li, and H. Zhang, “A novel facial feature point localization algorithm using probabilistic-like output,” in *Asian Conference on Computer Vision*, Jeju Island, Korea, 2004.
- [114] S. Shan, Y. Chang, W. Gao, B. Cao, and P. Yang, “Curse of mis-alignment in face recognition: Problem and a novel mis-alignment learning solution,” in *IEEE International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, 2004, pp. 314–320.
- [115] O. Boiman and M. Irani, “Similarity by composition,” in *Advances in Neural Information Processing Systems*, Cambridge, MA, USA, 2006, pp. 177–184.
- [116] J. Munkres, “Algorithms for the assignment and transportation problems,” *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, Mar 1957.
- [117] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 1470–1477.
- [118] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan 1967.
- [119] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Krger, and R. Chellappa, “Identification of humans using gait,” *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1163–1173, Sep 2004.
- [120] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*, Morgan & Claypool, California, USA, 2009.
- [121] J. Bonnans, J. Gilbert, C. Lemarchal, and Sagastizbal, *Numerical optimization, theoretical and numerical aspects (Second edition)*, Springer, New York, USA, 2006.
- [122] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

REFERENCES

- [123] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, NY, USA, 1987.
- [124] X. Zhou, X. Zhuang, S. Yan, S. F. Chang, M. Hasegawa-Johnson, and T. S. Huang, “Sift-bag kernel for video event analysis,” in *ACM International Conference on Multimedia*, Vancouver, British Columbia, Canada, 2008, pp. 229–238.
- [125] Y. Huang, D. Xu, and T. J. Cham, “Face and human gait recognition using image-to-class distance,” *IEEE Transactions on Circuits Systems for Video Technology*, vol. 20, no. 3, pp. 431 – 438, Mar 2010.
- [126] F. Nie, S. Xiang, Y. Jia, and C. Zhang, “Semi-supervised orthogonal discriminant analysis via label propagation,” *Pattern Recognition*, vol. 42, no. 11, pp. 2615 – 2627, 2009.
- [127] J. W. Right, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation, and its online supplementary material,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–226, 2009.

List of Publications

Published Papers

Journal

1. **Yi Huang**, Michael S. Brown and Dong Xu, “User-Assisted Ink-Bleed Reduction”, IEEE Transactions on Image Processing (T-IP), 2010.
2. **Yi Huang**, Dong Xu, and Tat-Jen Cham, “Face and human gait recognition using image-to-class distance”, IEEE Transactions on Circuits Systems for Video Technology (T-CSVT), vol. 20, pp. 431 C 438, Mar 2010.
3. **Yi Huang**, Dong Xu and Feiping Nie, “Patch Distribution Compatible Semi-Supervised Dimension Reduction for Face and Human Gait Recognition”, Circuits Systems for Video Technology (T-CSVT) **Accepted**

Conference

1. **Yi Huang**, Michael S. Brown and Dong Xu, “A Framework for Reducing Ink-Bleed in Old Documents”, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2008. Anchorage, Alaska, USA, Jun 2008, pp. 1 - 7.
2. **Yi Huang** and Michael S. Brown “User-Assisted Ink-Bleed Correction for Hand-written Documents”, IEEE and ACM Joint Conference on Digital Libraries (JCDL), Pittsburgh, Pennsylvania, USA, Jun 2008, pp. 263 - 271.[**Best Student Paper Award**]

3. **Yi Huang**, Dong Xu and Feiping Nie “Regularized Trace Ratio Discriminant Analysis with Patch Distribution Feature for Human Gait Recognition”, IEEE International Conference on Image Processing (ICIP), Hong Kong, China, Sep 2010, pp. 2449 - 2452.

Submitted Papers

Journal

1. **Yi Huang**, Dong Xu and Feiping Nie, “Trace Ratio Based Flexible Semi-supervised Discriminant Analysis”, Submitted to IEEE Transactions on Neural Networks (TNN)