A Unified Framework for Subspace Based Face Recognition

By

Wang Xiaogang

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Master of Philosophy

in

Information Engineering

© The Chinese University of Hong Kong

June 2003

The Chinese University of Hong Kong holds the copyright of this thesis. Any person(s) intending to use a part or whole of the materials in the thesis in a proposed publication must seek copyright release from the Dean of the Graduate School.

香港中文大學
大學
統系館書圖
28 APR 2004
UNIVERSITY
LIBRARY SYSTEM
THE CHINESE UNIVERSITY OF HONG KONG

Abstract

The family of subspace recognition methods represents the state-of-the-art in face recognition. PCA, LDA and Bayesian analysis are three of the most representative subspace based face recognition approaches. In this thesis, we show that they can be unified under the same framework. The difference of face images can be modeled as three major components: intrinsic difference, transformation difference, and noise. A unified framework is then constructed by using the face difference model and a detailed subspace analysis on the three face difference components. We explain the inherent relationship among different subspace methods and their unique contributions to the extraction of discriminating information from the face difference. PCA and Bayes can be viewed as intermediate steps of LDA. However, conventional LDA fails to attain the best performance without significant changes in each individual step.

Starting from the framework, a unified subspace analysis is developed using PCA, Bayes, and LDA as three steps. A 3D parameter space is constructed using the three subspace dimensions as axes. Searching through this parameter space, we achieve better recognition performance than the standard subspace methods. Analyzing the special requirement in each step, the unified subspace analysis adopts different training data at different steps. It avoids the conflicts between the large class number and small sample size in face recognition.

The drawbacks of standard subspace methods can be well analyzed under this framework. Taking advantage of the unified subspace analysis, several other novel subspace based face recognition approaches have been developed in this thesis, including discriminant analysis in dual intrapersonal subspaces and eigentransformation.

When the transformation difference between face images is significant, it can no longer be modeled as a Gaussian distribution, and the difference of face images cannot even be modeled as the linear composition of the three components (intrinsic difference, transformation difference, and noise). In this case, the framework for subspace methods will not work well. It is also a critical problem for conventional PCA, Bayes, and LDA. We develop an eigentransformation

approach to transform different style of face images into the same modality, such that the transformation difference can be significantly reduced. In this thesis, this novel approach is applied to two particular applications: recognizing face photos using sketch drawings and hallucination. In face sketch recognition, by transforming a photo into a sketch, we reduce the difference between photo and sketch significantly. A Bayesian classifier is then used to recognize the probe sketch from the synthesized pseudo-sketch. We also successfully apply eigentransformation to face hallucination, i.e. rendering the high-resolution face image from the low-resolution one.

摘要

在當今人像識別技術的最新發展中，基于子空間的人像識別技術是一類重要的方法。主分量分析、線性判別式分析和貝葉斯分析是三種最具有代表性的基于子空間的人像識別方法。論文中，我們將證明它們可以統一在同一個理論框架中。人臉圖像的差異可以分解成三個主要成分：由於身份不同引起的固有差異，由於光纖、姿勢、表情不同引起的變換差異，以及噪聲。基于這個人臉差異模型，通過對人臉差異的三個主要成分進行詳細的子空間分析，我們爲基于子空間的人像識別技術建立了一個統一的理論框架。這一理論框架論證了三種不同的子空間人像識別方法的內在聯繫，在從人臉差異中提取分類信息時它們各自不同的作用。主份量分析和貝業斯分析可以看做線性判別式分析的中間步驟。然而，傳統的線性判別式分析因爲沒有認識到這一點，從而不能對每一個步驟加以改進，取得最佳的識別效果。

基于這一理論框架，利用主分量分析、貝業斯分析和線性判別式分析作爲三個子步驟，我們提出了子空間綜合分析算法。利用三個子空間的維數作爲軸向量，我們構造了一個三維參數空間。搜索這個參數空間，可以得到比傳統的子空間算法更好的識別效果。通過分析每個步驟的特定要求，子空間綜合分析算法在訓練中針對不同的步驟采用不同的訓練集。它成功的解決了人像識別中遇到的類別數目多，訓練樣本少的問題。

在這個理論框架下，我們可以清楚的分析傳統的基于子空間的人像識別方法存在的不足。除了子空間綜合分析算法外，論文中進而又提出了其它幾種新的基于子空間的人像識別算法，包括基于雙重類內變化子空間的線性判別式算法和基于主分量分析的子空間轉換算法。

有時由於外界因素的存在，人臉圖像會有較大的變換差異，不能用高斯分布進行近似。人臉的圖像差異甚至不能表達成三種成分(固有差異、變換差異、噪聲)的線性組合。基于子空間分析的理論框架就會失效。這也是傳統主份量分析、貝葉斯分析、和線性判別式分析所共同遇到的一個重要問

題。我們提出了一個基于主份量分析的空間變換方法，可以將不同“風格”的人臉圖像轉換爲同一“風格”，這樣變換差異就會被有效地減小。論文中，我們將這個新方法用于兩個特殊的應用：利用素描畫像識別照片和人臉圖像分辨率的增強。在畫像識別的應用中，我們將照片轉換成畫像，這樣就有效的減少了照片與畫像之間的差異。這個方法還可以用于從低分辨率的人臉圖像中恢復出高分辨率的人臉圖像。

Acknowledgments

Here I would like to acknowledge all the people who had assisted me during the past two years of my graduate studies at the Chinese University of Hong Kong. I am most grateful to my supervisor, Dr. Xiaoou Tang. All the research work in this thesis is completed under his professional and careful direction. He has proposed many important and valuable ideas and suggestions for my research, and helped me greatly improve the presentation of this thesis. I have learnt so much from him in the past two years. I am very fortunate to be able to complete my postgraduate study under his direction.

I would like to thank Prof. Jianzhuang Liu. He gave some very useful suggestions to my work. I also would like to thank all my partners in the multimedia laboratory, Feng Lin, Lifeng sha, Feng Zhao, Zhifeng Li, Bo Luo, Hua Shen, Tong Wang, and Dacheng Tao. We have lived and studied for two years. I always can get kindly help and encouragement from them.

I owe my sincere thanks to my parents, for their never fading love, care, understanding and encouragement.

Table of Contents

Abstract i

Acknowledgments v

Table of Contents vi

List of Figures viii

List of Tables x

Chapter 1 Introduction 1

1.1 Face recognition 1
1.2 Subspace based face recognition technique 2
1.3 Unified framework for subspace based face recognition 4
1.4 Discriminant analysis in dual intrapersonal subspaces 5
1.5 Face sketch recognition and hallucination 6
1.6 Organization of this thesis 7

Chapter 2 Review of Subspace Methods 8

2.1 PCA 8
2.2 LDA 9
2.3 Bayesian algorithm 12

Chapter 3 A Unified Framework 14

3.1 PCA eigenspace 16
3.2 Intrapersonal and extrapersonal subspaces 17
3.3 LDA subspace 18
3.4 Comparison of the three subspaces 19
3.5 L-ary versus binary classification 22
3.6 Unified subspace analysis 23
3.7 Discussion 26

Chapter 4 Experiments on Unified Subspace Analysis 28

4.1 Experiments on FERET database 28
4.1.1 PCA Experiment 28
4.1.2 Bayesian experiment 29
4.1.3 Bayesian analysis in reduced PCA subspace 30
4.1.4 Extract discriminant features from intrapersonal subspace 33
4.1.5 Subspace analysis using different training sets 34
4.2 Experiments on the AR face database 36
4.2.1 Experiments on PCA, LDA and Bayes 37
4.2.2 Evaluate the Bayesian algorithm for different transformation 38

Chapter 5 Discriminant Analysis in Dual Subspaces 41

5.1 Review of LDA in the null space of S_w and direct LDA 42
5.1.1 LDA in the null space of S_w 42

5.1.2 Direct LDA 43
5.1.3 Discussion 44
5.2 Discriminant analysis in dual intrapersonal subspaces 45
5.3 Experiment 50
5.3.1 Experiment on FERET face database 50
5.3.2 Experiment on the XM2VTS database 53

Chapter 6 Eigentransformation: Subspace Transform 54

6.1 Face sketch recognition 54
6.1.1 Eigentransformation 56
6.1.2 Sketch synthesis 59
6.1.3 Face sketch recognition 61
6.1.4 Experiment 63
6.2 Face hallucination 69
6.2.1 Multiresolution analysis 71
6.2.2 Eigentransformation for hallucination 72
6.2.3 Discussion 75
6.2.4 Experiment 77
6.3 Discussion 83

Chapter 7 Conclusion 85

Publication List of This Thesis 87

Bibliography 88

List of Figures

Figure 1-1 Examples of face appearance changes under different disturbing factors. 2
Figure 2-1 Eigenvectors and eigenvalues for a 2D distribution. 9
Figure 2-2 Compare PCA and LDA for a two-class problem. 10
Figure 2-3 Example of simultaneous diagonalization of S_w and S_b. 11
Figure 2-4 Decompose image space into principal subspace F and complementary subspace $\overline{F}$. 13
Figure 3-1 Diagram of the unified framework for subspace based face recognition. 16
Figure 3-2 Energy distribution of three components $\widetilde{I}$, $\widetilde{T}$, and $\widetilde{N}$ on eigenvectors in three subspaces. 20
Figure 3-3 Relationship of the PCA, Bayes, and LDA subspaces. 21
Figure 3-4 Use average intrapersonal variation distribution to approximate that for each individual class. 23
Figure 3-5 3D parameter space. 25
Figure 4-1 Example of normalized face image. 29
Figure 4-2 Recognition accuracy of PCA on the FERET database. 29
Figure 4-3 Recognition accuracy of the Bayesian algorithm on the FERET database. 30
Figure 4-4 Accuracy surface for the Bayesian analysis in the PCA subspace. 32
Figure 4-5 Highest accuracy of the Bayesian analysis in each PCA subspace. 33
Figure 4-6 Accuracies using different number of discriminant features extracted from intrapersonal subspace. 34
Figure 4-8 Subspace analysis for different training sets. 35
Figure 4.9 Samples for the seven transformations in AR database. 36
Figure 4-10 First five eigenfaces for different subspaces. 37
Figure 4-11 Accuracies of direction correlation and Bayesian algorithm for different transformations. 40
Figure 5-1 Direct LDA for a two-class problem. 44
Figure 5-2 Analysis different LDA approaches. 45
Figure 5-3 Training discriminant vectors in dual intrapersonal subspaces. 47
Figure 5-4 Recognition accuracy comparison of the new method with Bayesian face recognition. 52
Figure 5-5 Accumulative scores for the new method, Bayes (Mahalanobis distance), Fisherface, LDA in null space, and direct LDA. 52
Figure 5-6 Recognition accuracies of Fisherface, LDA in null space, direct LDA, and the new method using different number of persons for training on the FERET database. 53
Figure 6-1 Examples of photo-sketch pairs. 55
Figure 6-2 Eigentransformation procedure. 57
Figure 6-3 Eigentransformation with the assumption that the transformation between photo and sketch is linear. 58
Figure 6-4 Framework of the face sketch synthesis system. 61
Figure 6-5 Face sketch recognition using eigentransformation and the Bayesian classifier. 62
Figure 6-6 Facial sketch synthesis based on full face. 64
Figure 6-7 Generate photo from the input sketch. 65
Figure 6-8 Comparison of the direct eigentransformation (first row) with separate transformation on texture and shape (second row). 66

Figure 6-9 Comparison of accumulative match score between our automatic recognition method and human performance. 69
Figure 6-10 Multiresolution analysis in spatial domain. 70
Figure 6-11 System diagram using eigentransformation for hallucination. 73
Figure 6-12 Eigenfaces sorted by eigenvalues. e_i is the ith eigenface. 76
Figure 6-13 Extract facial information in the PCA space of low-resolution face image. 76
Figure 6-14 Hallucinated face images by eigentransformation. 78
Figure 6-15 Hallucinated face images using different resolutions as the input. 81
Figure 6-16 RMS error per pixel in intensity using Cubic spline interpolation and hallucination by eigentransformation. 81
Figure 6-17 Adding constrain to the principal components of the hallucinated face images. 82
Figure 6-18 Recognition accuracy using low-resolution face images and hallucinated face images based on XM2VTS database. 82
Figure 6-19 Hallucinating face with additive zero mean, Gaussian noise. 83

List of Tables

Table 3-1 Behavior of subspace on characterizing the face image difference. 22
Table 4-1 Recognition accuracy of Bayesian analysis in reduced PCA subspace (%). 32
Table 4-2 Seven transformations for each individual class in each session from AR database. .36
Table 4-3 Recognition result of PCA, LDA, and Bayes (ML) on AR databse. 38
Table 4-4 Recognition accuracies of LDA on AR database using different feature number..... 38
Table 4-5 The testing and training sets for different transformations. 39
Table 4-6 Recognition results of direct correlation and Bayesian algorithm for different transformations..... 39
Table 5-1 Recognition accuracies of Fisherface, LDA in null space, direct LDA, and the new method using different numbers of persons for training on the FERET database. 52
Table 5-2 Face recognition accuracies on the four experimental trials of the data set from the XM2VTS database. 53
Table 6-1 Recognition accuracies using different features and classifiers (%). 67
Table 6-2 Acumulative match score for eigenface, EGM, and the novel method (%)..... 67

Chapter 1

Introduction

1.1 Face recognition

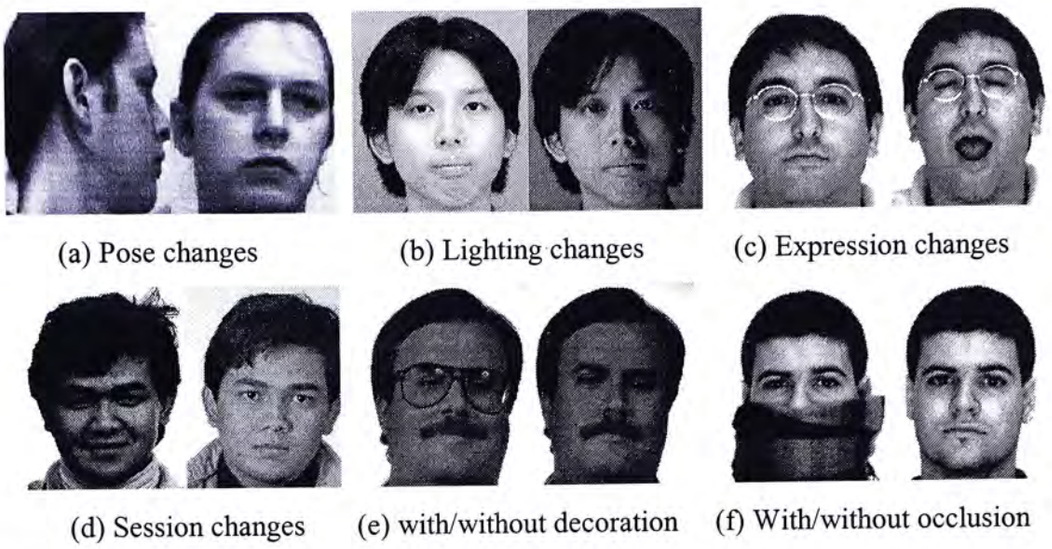
Automatic face recognition as an important Biometrics technique has drawn more and more attention in recent years. Comparing to other individual identification techniques, face recognition is more convenient under real world operation conditions, since it does not require those being watched to cooperate. It has been widely used in law enforcement identification, banking and security system access authentication, and anti-terrorist video surveillance, etc.

A general statement of automatic face recognition is described in [83]: "Given still or video images of a scene, identify or verify one or more persons in the scene using a stored database of faces." Therefore, the two main tasks of face recognition are outlined as two categories,

- **Face identification**: Given an unknown face as input, the system determines the identity through a one-to-many matching with all the known individuals in the database. The system usually returns the N most similar reference faces to the test face.
- **Face verification**: The system confirms or rejects the claimed identity of the input face.

The main challenge for face identification and verification is that even though human faces share similar features, face images belonging to the same individual may have very different under different conditions. The factors affecting face appearance include,

- Pose
- Lighting
- Expression
- Session changes
- With/or without decoration and occlusion



(a) Pose changes (b) Lighting changes (c) Expression changes

(d) Session changes (e) with/without decoration (f) With/without occlusion

Figure 1-1 Examples of face appearance changes under different disturbing factors.

Some example face images affected by these factors are shown in Figure 1. In face recognition study, it is critical to distinguish whether the appearance variation is caused by face identity or other disturbing factors.

1.2 Subspace based face recognition technique

Many face recognition techniques have been developed over the past thirty years. A detailed survey can be found in [62][83]. Most of the face recognition techniques can be categorized into two classes: feature-based methods and appearance-based methods. Feature-based methods extract the geometrical relationship and other parameters of face features for matching. Appearance-based approaches view a 2D image as a vector in the high dimensional image space. A suitable metric is then used for face matching in the image space or its subspace. A comparative study by Brunelli and Poggio [61] shows that appearance-based techniques have superior performance than feature-based techniques.

Face image appearance began to be used for recognition in the early 1980s. Baron [65] develops a direct appearance-based matching procedure: correlation between 2D raw images. Because of the high dimensionality of the raw image, the direct correlation is expensive to compute. In order to reduce the dimensionality of the original face image, many subspace methods have been developed to extract more compact features for face recognition. The eigenface method (PCA) developed by Turk and Pentlend [47][51][52] is a major breakthrough for the appearance-based techniques. The method uses the Karhunen-Loeve Transform to

produce a most expressive subspace for face representation and recognition. Inspired by the eigenface approach, several appearance-based subspace methods have been developed. LDA or Fisher Face [58][84][35][85][22] is an example of the most discriminating subspace methods. Linear Discriminant Analysis (LDA) is adopted to seek a set of features best separating face classes. Another important subspace method is the Bayesian algorithm using probabilistic subspace proposed by Moghaddam [8][9][10][11]. Different from other subspace techniques, which classify the test face image into L classes for L individuals, the Bayesian algorithm casts the face recognition task into a binary pattern classification problem with each of the two classes, intrapersonal variation and extrapersonal variation, modeled by a Gaussian distribution.

Many other subspace methods are more or less modification or extension of the above three methods. Pentlend et. al. [5] extend the eigenface method to view-based and modular eigenspaces. Craw et. al. [29] normalize the face image to a shape-free vector based on 34 fiducial points as the preprocessing for eigenface techniques. Independent Component Analysis (ICA) [54][55][48], nonlinear PCA (NLPCA) [41], Kernel PCA (KPCA) [13][53] are all the generalizations of PCA to address higher order statistical dependencies. Kernel-based Fisher Discriminant Analysis (KFDA) [23][60] extracts nonlinear discriminanting features. "Evolutionary Pursuit" [17] searches for the optimal basis in the whitened PCA space. Coarse-to-fine hierarchical discriminating subspaces are implemented by applying PCA and LDA projection recursively [80][21]. To improve the generalization ability of LDA on different data sets, several modifications are proposed, such as the Enhanced FLD model [16], LDA mixture model [25], and direct LDA [28] etc.

In addition to processing original image directly, subspace methods can also model other features, such as shape and wavelet features. Cootes and Taylor developed Active Appearance Model (AAM) [75][74][1] to explicitly model both shape and texture. Liu and Wechsler apply Enhanced Fisher Classifier on face recognition based on integrated shape and texture [18], and on Gabor features [19]. PCA and LDA have also been integrated with Fourier and wavelet descriptors [14][20].

1.3 Unified framework for subspace based face recognition

In this thesis, we develop a unified framework to study the three subspace face recognition methods: PCA, LDA and the Bayesian algorithm. As discussed earlier, the three methods represent three major approaches for subspace based face recognition. PCA has become an evaluation benchmark for face recognition. Both LDA and Bayesian algorithms have achieved superior performance in the FERET competition compared to other methods [40][41]. A unified framework on the three methods will greatly help to understand the family of subspace methods for further improvement of the methods.

The face difference between two face images can be modeled as three major components: intrinsic difference $\tilde{I}$ caused by face identity, transformation difference $\tilde{T}$ caused by pose, lighting, and expression changes etc., and noise $\tilde{N}$. A unified framework is then constructed by using the face difference model and a detailed subspace analysis on the three face difference components. PCA, Bayes, and LDA are initially developed under different consideration, and seem quite different from each other on the surface. Using this framework we explain the inherent relationship among the three different subspace methods and their unique contributions to the extraction of discriminating information from the face difference. PCA and Bayes can be viewed as intermediate steps of LDA. PCA reduces the noise. Bayes reduces the transformation difference, but may lead to the increase of the noise level. Based on PCA and Bayes, LDA further reduce the noise and compact the intrinsic difference to a small number of features. However, conventional LDA cannot attain the best performance without improving each individual step.

Starting from the framework, a unified subspace analysis is proposed using PCA, Bayes, and LDA as three steps [87]. It is pointed out that the subspace dimension of each method can affect the recognition performance. It is a trade-off on how much noise and transformation difference are excluded, and how much intrinsic difference is retained. This eventually leads to the construction of a 3D parameter space that uses the three subspace dimensions as axes. Searching through this parameter space, we achieve better recognition performance than the standard subspace methods, which are all constrained on local areas of the parameter space. Analyzing the special requirement of each step, our unified

subspace analysis adopts different training data at different steps of training process. Using the training samples outside gallery, the PCA subspace and intrapersonal subspace are effectively improved to reduce the noise and transformation difference. Only the class centers of the individuals in the gallery are used in the final step of discriminant analysis, so that the extracted discrimiant features are specially tuned for the individual in the gallery. This helps to avoid the conflict between the large class number and small sample size in face recognition.

The disadvantages of conventional subspace face recognition methods can be well understood under this framework. Starting from the framework, several other improvements to the subspace based face recognition are also proposed in this thesis.

1.4 Discriminant analysis in dual intrapersonal subspaces

The framework demonstrates that the high dimensional image space can be decomposed into intrapersonal principal subspace and its complementary subspace. Both subspaces contain discriminative information useful for recognition. Conventional LDA approaches, such as Fisherface, LDA in null space, and direct LDA, only apply discriminant analysis in one subspace, thus discards some discriminative information in the other subspace. The Bayesian algorithm makes use of the features in two subspaces, however it does not further apply discriminant analysis on the class centers. Moreover, in the Bayesian algorithm, computing the component in the intrapersonal complementary subspace is expensive since it does not compact the discriminative features to improve the recognition efficiency. Since Bayes can be viewed as the intermediate step of LDA, integrating the advantages of the two approaches, a novel face recognition approach is proposed to apply discriminant analysis in dual intrapersonal subspaces, and combine the two parts of discriminative features under a probabilistic model. It outperforms the Bayesian and LDA approaches in both recognition accuracy and computational efficiency.

1.5 Face sketch recognition and hallucination

In this thesis, we also study the subspace methods on two particular applications: face sketch recognition [86][88]and hallucination [90]. An important application of face recognition is to assist law enforcement. However, in most cases, the photo image of a suspect is not available. The best substitute is often a sketch drawing based on the recollection of an eyewitness. Therefore, automatically searching through a photo database using a sketch drawing is very useful. It will not only help the police to locate a group of potential suspects, but may also help the witness and the artist to modify the sketch drawing of the suspect interactively based on the similar photos retrieved.

However, due to the great difference between sketches and photos, and the unknown psychological mechanism of sketch generation, face sketch recognition is much more difficult than the normal face recognition based on photo image. Directly applying subspace framework is not practical. In this case, the transformation difference is too large to be modeled as Gaussian distribution, because photo and sketch are in different modalities. The face difference even cannot be modeled as linear composition of the three components (intrinsic difference, transformation difference, and noise). In this thesis, we develop an eigentransformation approach to transform a photo into a sketch, such that the difference between photo and sketch is significantly reduced. A Bayesian classifier is then used to recognize the probing sketch from the synthesized pseudo-sketch.

The eigentransformation algorithm also can be applied to face image hallucination, rendering a high-resolution face image from a low-resolution one. In video surveillance, the faces of interest are often in small size because of the large distance between the camera and the objects. Image resolution becomes an important factor affecting face recognition performance. Since many detail facial features are lost in the low-resolution face images, the faces are often indiscernible. Our hallucination method is not only much helpful for recognition by human, but also make the automatic recognition procedure easier, since it emphasizes the face difference by adding more high frequency details.

1.6 Organization of this thesis

The thesis is organized as following. In Chapter 2, we review three representative subspace face recognition methods, PCA, LDA, and the Bayesian algorithm. In Chapter 3, the three subspace methods are unified under a novel framework for subspace based face recognition, and a unified subspace analysis is proposed. Experimental analysis on the framework is given in Chapter 4. Chapter 5 develops a novel face recognition approach applying dsicriminant analysis in dual intrapersonal subspaces. The eigentransformation algorithm is proposed and applied to face sketch recognition and hallucination in Chapter 6. Chapter 7 draws the conclusion of this thesis.

Chapter 2

Review of Subspace Methods

In appearance-based approaches, a 2D face image is viewed as a vector in the image space. We formulate the face recognition problem as the following. A set of sample face images $\{\bar{x}_i\}$ can be represented as a N by M matrix $X = [\bar{x}_1, \ldots, \bar{x}_M]$, where N is the number of pixels in the images and M is the number of samples. Each face image $\bar{x}_i$ belongs to one of the L individual classes $\{X_1, \ldots, X_L\}$, with $\ell(\bar{x}_i)$ as the class label of $\bar{x}_i$. When a test image $\bar{T}$ is the input, the face recognition task is to find its class label in the database. Based on this formulation, a short review for the three subspace approaches is given in this section.

2.1 PCA

The PCA method uses the Karhunen-Loeve Transform for the representation and recognition of faces. A set of eigenvectors, also called eigenfaces, spans the subspace (eigenspace) of the image space. Eigenfaces are typically computed from the eigenvectors of sample covariance matrix C,

$$C = \sum_{i=1}^{M} (\bar{x}_i - \bar{m})(\bar{x}_i - \bar{m})^T , \tag{2-1}$$

where $\bar{m}$ is the mean face computed from the sample set

$$\bar{m} = \frac{1}{M} \sum_{i=1}^{M} \bar{x}_i . \tag{2-2}$$

The eigenspace U is spanned by the K eigenfaces with the largest eigenvalues, $U = [\bar{u}_1, \ldots \bar{u}_K]$. As shown in Figure 2-1, eigenvalues characterizes the variation of face set on the eigenfaces. The K eigenfaces with the largest eigenvalues capture the most variation of human face. For a face image $\bar{x}$, it is removed of the mean face and projected to eigenspace defined by U to get the weight vector,

$$\bar{w} = U^T (\bar{x} - \bar{m}). \tag{2-3}$$

$\bar{w}$ can be used for face representation and recognition. $\bar{x}$ can be optimally reconstructed from $\bar{w}$ using only K features with the minimum reconstruction error,

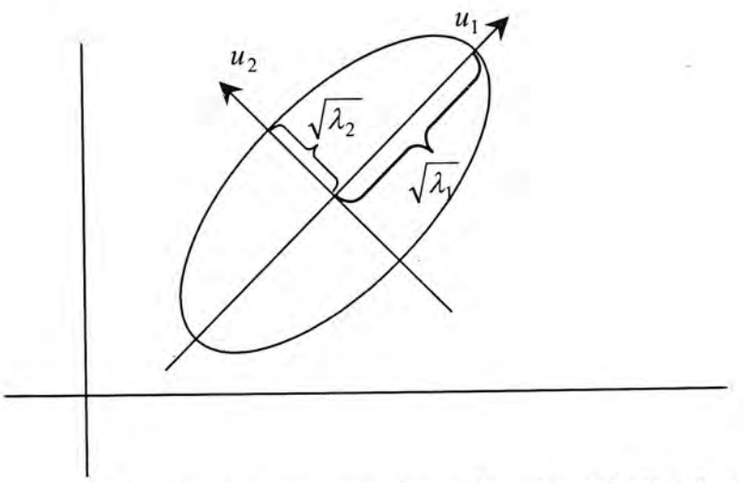


Figure 2-1 Eigenvectors and eigenvalues for a 2D distribution.

u_i is the eigenvector and λ_i is the eigenvalue.

$$\vec{x}_r = U\vec{w} + \vec{m} \tag{2-4}$$

In the recognition process, the prototype $\vec{P}$ for each individual class in the database and the test image $\vec{T}$ to be classified are projected onto the eigenspace to get the prototype weight vector $\vec{w}_P$ and test weight vector $\vec{w}_T$,

$$\vec{w}_p = U^T\left(\vec{P} - \vec{m}\right). \tag{2-5}$$

$$\vec{w}_T = U^T\left(\vec{T} - \vec{m}\right). \tag{2-6}$$

The face class is found to minimize the distance

$$\varepsilon = \| \vec{w}_T - \vec{w}_p \| . \tag{2-7}$$

2.2 LDA

The features extracted by PCA method are best for face representation, but not optimal for face classification. Different from PCA, the LDA method tries to find the subspace that best discriminates different face classes. A comparison of PCA and LDA is shown in Figure 2-2. For a two-class problem, a projection with the largest total scatter is not necessarily good for classification. The LDA projections are achieved by maximizing the between-class scatter matrix S_b, while minimizing the within-class scatter matrix S_w. S_w and S_b are defined as

$$S_w = \sum_{i=1}^{L} \sum_{\vec{x}_k \in X_i} (\vec{x}_k - \vec{m}_i)(\vec{x}_k - \vec{m}_i)^T , \tag{2-8}$$

$$S_b = \sum_{i=1}^{L} n_i (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T , \tag{2-9}$$

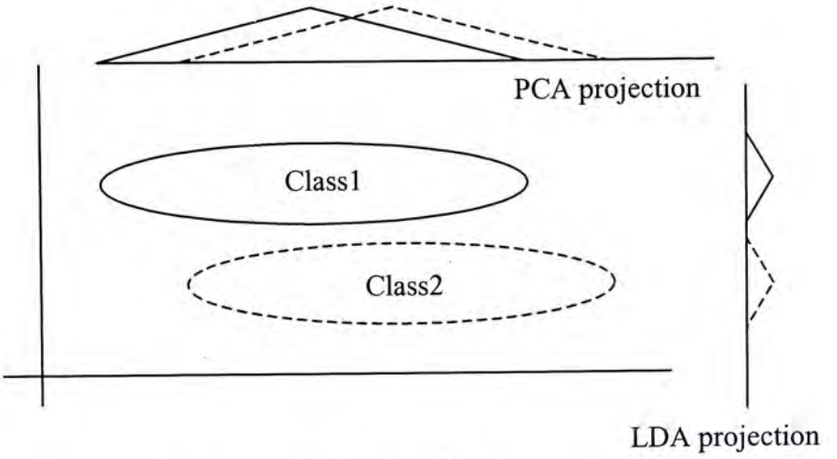


Figure 2-2 Compare PCA and LDA for a two-class problem.

PCA chooses the projection with the maximum total scatter. LDA chooses the projection with minimum within-class variation and maximum between-class variation. Usually, LDA outperforms PCA in classification

where $\bar{m}_i$ is the mean face for the individual class X_i, and n_i is the number of samples in class X_i.

The subspace for LDA is spanned by a set of vectors $W = [\bar{w}_1, \ldots, \bar{w}_{L-1}]$, satisfying

$$W = \arg\max \left| \frac{W^T S_b W}{W^T S_w W} \right|, \tag{2-10}$$

W can therefore be constructed by the eigenvectors of $S_w^{-1} S_b$. Computing the eigenvectors of $S_w^{-1} S_b$ is equivalent to simultaneous diagonalization of S_w and S_b [36]. First S_w is whitened by.

$$\Theta^{-1/2} \Phi^T S_w \Phi \Theta^{-1/2} = I, \tag{2-11}$$

where Φ and Θ are the eigenvector matrix and eigenvalue matrix of S_w. Second, apply PCA on class centers of the transformed data. To do this, we project the class centers onto $\Theta^{-1/2} \Phi^T$, thus the between-class matrix is transformed to K_b as,

$$K_b = \Theta^{-1/2} \Phi^T S_b \Phi \Theta^{-1/2}. \tag{2-12}$$

After computing the eigenvector matrix Ψ and eigenvalue matrix Λ of K_b, the overall projection vectors of LDA can be defined as

$$W = \Phi \Theta^{-1/2} \Psi. \tag{2-13}$$

Since matrix I is invariant under transformation Ψ,

$$\Psi^T I \Psi = \Psi^T \Psi = I, \tag{2-14}$$

and

$$\Psi^T K_b \Psi = \Lambda, \tag{2-15}$$

we have

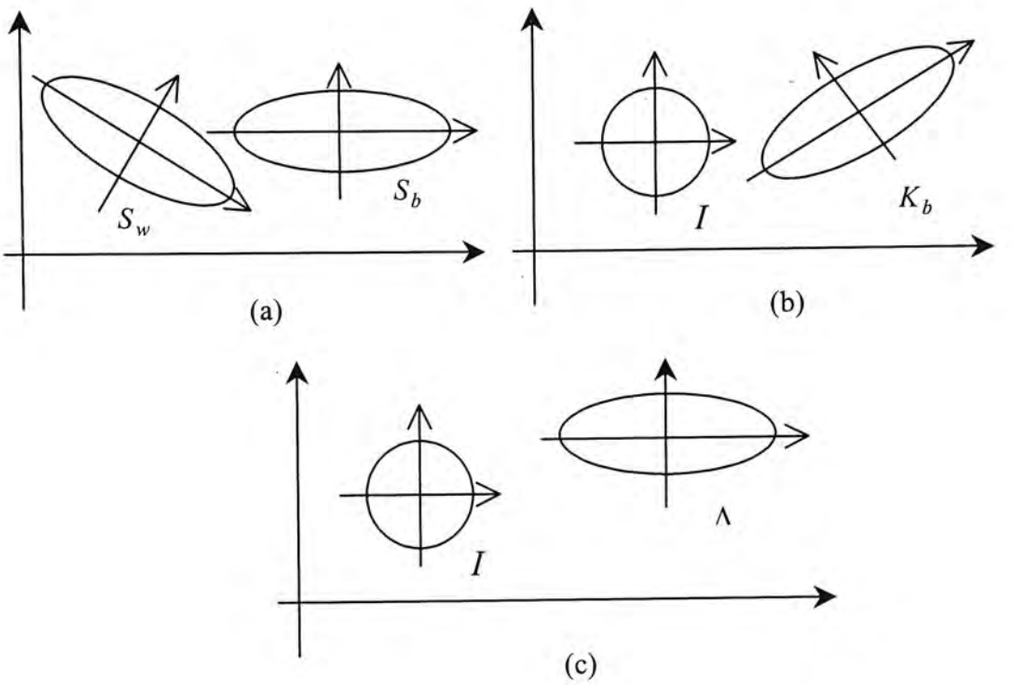


Figure 2-3 Example of simultaneous diagonalization of S_w and S_b.

(a): S_w and S_b ; (b) After whitening by the eigenvectors and eigenvalues of S_w, S_w is transformed to I and S_b is transformed to K_b ; (c): Projecting to the eigenvectors of K_b, S_w is transformed to I, and S_b is transformed to Λ

$$W^T S_w W = I, \tag{2-16}$$

$$W^T S_b W = \Lambda. \tag{2-17}$$

As shown in [28], W is the eigenvector matrix of $S_w^{-1} S_b$. A two-dimensional example of this process is shown in Figure 2-3.

For recognition, the linear discriminant function for the class prototype $\bar{P}$ and test image $\bar{T}$ is thus computed as,

$$d(\bar{T}) = W^T(\bar{T} - \bar{P}). \tag{2-18}$$

The face class is chosen to minimize $\| d \|$.

Usually the dimension of face vector is far larger than the training samples ($N>>M$). Since the rank of S_w is at most M-L, when dealing with the high dimensional face data, S_w will become singular, and the LDA vectors are difficult to compute. To avoid degeneration of S_w, most LDA methods first reduce the data dimensionality by PCA, then apply discriminant analysis in the reduced PCA space. In Fisherface [58], the dimension of PCA space is fixed as M-L, and L-1 LDA features are extracted for recognition.

2.3 Bayesian algorithm

Different from other subspace methods, which classify the test face image $\bar{T}$ into L classes for L individuals, the Bayesian algorithm classifies the face intensity difference $\Delta = \bar{T} - \bar{P}$ as intrapersonal variation (Ω_I) for the same individual and extrapersonal variation (Ω_E) for different individuals. The MAP similarity between two images is defined as the intrapersonal a posterior probability,

$$S(I_1, I_2) = P(\Omega_I \mid \Delta)$$

$$= \frac{P(\Delta \mid \Omega_I) P(\Omega_I)}{P(\Delta \mid \Omega_I) P(\Omega_I) + P(\Delta \mid \Omega_E) P(\Omega_E)}. \tag{2-19}$$

The more similar two face images, the larger $S(I_1, I_2)$. Because of the high dimensionality, $P(\Delta \mid \Omega_I)$ and $P(\Delta \mid \Omega_E)$ are difficult to be estimated directly from the training set. So subspace estimation is used instead. Both $P(\Delta \mid \Omega_I)$ and $P(\Delta \mid \Omega_E)$ are modeled as Gaussian distribution. To estimate $P(\Delta \mid \Omega_I)$, as illustrated in Figure 2-4, PCA on the intrapersonal difference set $\{\Delta \mid \Delta \in \Omega_I\}$ decomposes the image difference space into principal subspace F also called intrapersonal eigenspace, spanned by K the largest intrapersonal eigenvectors, and its orthogonal complementary space $\overline{F}$, with the dimension N-K. The likehood can be estimated as the product of two independent marginal Gaussian densities in F and $\overline{F}$,

$$\hat{P}(\Delta \mid \Omega_I) = \left[\frac{\exp\left(-\frac{1}{2} d_F(\Delta)\right)}{(2\pi)^{K/2} \prod_{i=1}^{K} \lambda_i^{1/2}} \right] \left[\frac{\exp\left(-\varepsilon^2(\Delta)/2\rho\right)}{(2\pi\rho)^{(N-K)/2}} \right] = \frac{\exp\left[-\frac{1}{2}\left(d_F(\Delta) + \varepsilon^2(\Delta)/\rho\right)\right]}{\left[(2\pi)^{K/2} \prod_{i=1}^{K} \lambda_i^{1/2}\right]\left[(2\pi\rho)^{(N-K)/2}\right]}. \tag{2-20}$$

$\hat{P}(\Delta|\Omega_I)$ is an estimation to $P(\Delta|\Omega_I)$. In Eq. (2-20), $d_F(\Delta)$ is a Mahalanobis distance in F, referred as "distance-in-feature-space" (DIFS),

$$d_F(\Delta) = \sum_{i=1}^{K} \frac{y_i^2}{\lambda_i}, \tag{2-21}$$

where y_i is the principal component of Δ projecting to the ith intrapersonal eigenvector, and λ_i is the corresponding eigenvalue. $\varepsilon^2(\Delta)$ is defined as "distance-from-feature-space" (DFFS), which is equivalent to PCA residual error in $\overline{F}$. ρ is the average eigenvalue in $\overline{F}$,

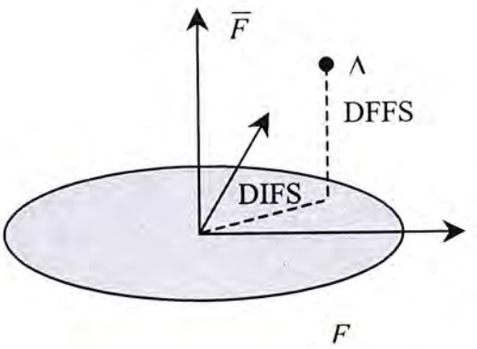


Figure 2-4 Decompose image space into principal subspace F and complementary subspace $\overline{F}$. DIFS and DFFS are computed to estimate the likehood.

$$\rho = \frac{1}{N-K}\sum_{i=K+1}^{N}\lambda_i \,. \tag{2-22}$$

$P(\Delta \mid \Omega_E)$ can be estimated in a similar way. The principal subspace computed from the set $\{\Delta \mid \Delta \in \Omega_E\}$ is called extrapersonal eigenspace.

In the matching process, the difference between the test image and the prototype in the database is first projected onto the intrapersonal and extrapersonal subspaces to estimate the Gaussian likehood by (2-20). The likehoods are combined in (2-19) and the class is found by comparing the similarity measure.

An alternative maximum likehood (ML) measure is proved to be simpler but almost as effective as the above MAP measure [9]. It uses the intrapersonal likehood alone as the similarity measure,

$$S'(\Delta) = P(\Delta \mid \Omega_I)\,. \tag{2-23}$$

Chapter 3

A Unified Framework

The three methods reviewed in Section 2 are developed under different circumstances by different researchers with different and specific considerations. Therefore, on the surface, these methods seem quite different from each other. In this study, instead of conducting a simple experimental comparison of the three methods, we formulate an in-depth subspace analysis to construct a unified framework for the three methods. Under this framework, we study the inherent connections of the three methods in order to discover the reason behind the different performances of each method under different circumstances. This is critically important for future development of new algorithms.

To construct the framework, let us first look at the matching criterions and focus on the difference $\Delta = \bar{T} - \bar{P}$ between the test image $\bar{T}$ and the prototype $\bar{P}$. The matching criterion for PCA in (2-7) can be rewritten as

$$\begin{aligned} \varepsilon_{PCA} &= \| U^T(\bar{T} - \bar{m}) - U^T(\bar{P} - \bar{m}) \| \\ &= \| U^T(\bar{T} - \bar{P}) \| \\ &= \| U^T \Delta \| . \end{aligned} \tag{3-1}$$

For the LDA method, according to (2-18), the linear discriminant function can also be expressed in terms of Δ,

$$\varepsilon_{LDA} = \| W^T \Delta \| . \tag{3-2}$$

Finally, the probabilistic measure in the Bayesian analysis can be translated into a distance measure. In recognition, all the parameters in Eq. (2-20) are constant except $d_F(\Delta)$ and $\varepsilon^2(\Delta)$. So the ML measure is equivalent to evaluating the distance,

$$D_I = d_F(\Delta) + \varepsilon^2(\Delta)/\rho . \tag{3-3}$$

Another distance D_E can be defined in the same way in the extrapersonal subspace. From (2-19), the MAP measure can be reformulated as

$$S(I_1, I_2) = \frac{P(\Omega_I)}{P(\Omega_I) + \frac{P(\Delta \mid \Omega_E)}{P(\Delta \mid \Omega_I)} P(\Omega_E)} . \tag{3-4}$$

Since $P(\Omega_I)$ and $P(\Omega_E)$ are fixed in matching procedure, the MAP measure only depends on the ratio of the two likehoods $P(\Delta \mid \Omega_I)$ and $P(\Delta \mid \Omega_E)$,

$$\frac{P(\Delta|\Omega_I)}{P(\Delta \mid \Omega_E)} \propto \exp\left[-\frac{1}{2}(D_I - D_F)\right]. \tag{3-5}$$

Therefore it can be further simplified to

$$\varepsilon_{Bayes} = D_I - D_E \tag{3-6}$$

From (3-1), (3-2), and (3-3), we can see that the recognition process of the three methods can be described by the same framework as shown in Fig. 3-1. When a test face image $\bar{T}$ is the input, we compute the difference Δ between $\bar{T}$ and each class prototype $\bar{P}$. The difference Δ is then projected onto an image subspace to compute the feature vector $\bar{V}_D$. Finally based on the feature vector and the specific distance metric, Δ is classified as either intrapersonal variation or extrapersonal variation.

The two central components of this framework are the image difference Δ and the subspace onto which Δ is projected. We model the image difference Δ by three key components: intrinsic difference ($\tilde{I}$) discriminating face identity; transformation difference ($\tilde{T}$), arising from all kinds of transformations, such as expression, illumination, and pose changes; noise ($\tilde{N}$), which randomly distributes in the face images.

The intrapersonal variation Ω_I is composed of $\tilde{T}$ and $\tilde{N}$, since it comes from the same individual. Extrapersonal variation is not equivalent to $\tilde{I}$. In Ω_E, $\tilde{I}$, $\tilde{T}$, and $\tilde{N}$ are coupled together, since $\tilde{T}$ and $\tilde{N}$ cannot be canceled when computing the difference of the images of two individuals. Therefore, we have,

$$\Omega_I = \tilde{T} + \tilde{N}, \tag{3-7}$$

$$\Omega_E = \tilde{I} + \tilde{T} + \tilde{N}. \tag{3-8}$$

$\tilde{I}$ contains the features discriminating different classes. $\tilde{T}$ and $\tilde{N}$ are the two components that deteriorate the recognition performance. Normally, $\tilde{N}$ is of small energy. The main difficulty for face recognition comes from transformations, which can change the face image appearance substantially. Under a large transformation, $\tilde{T}$ can potentially be greater than $\tilde{I}$ [91]. A successful face recognition algorithm should be able to reduce the energy of $\tilde{T}$ as much as possible without sacrificing much of $\tilde{I}$. To improve recognition efficiency, it is

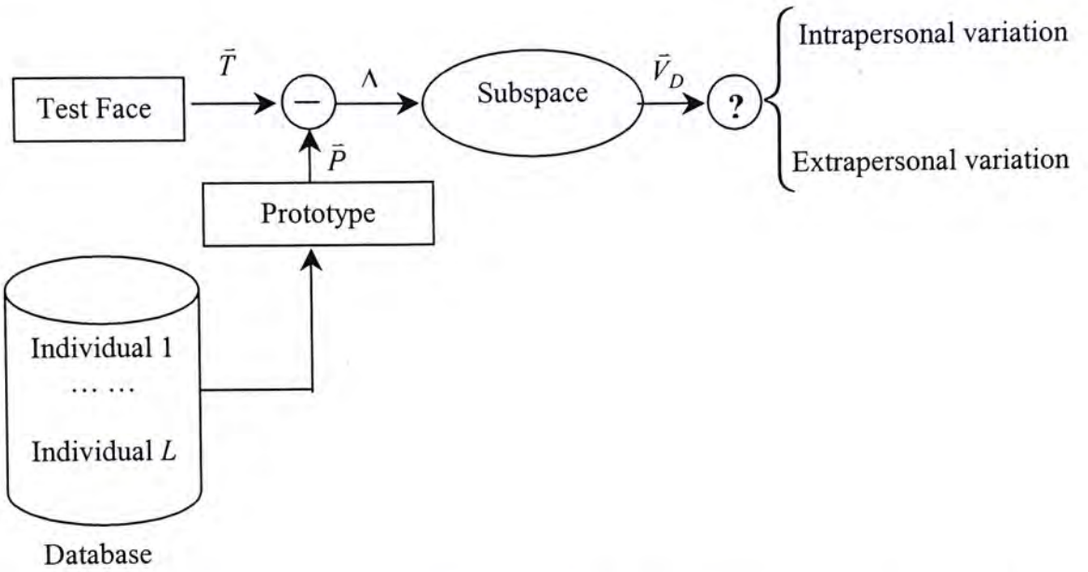


Figure 3-1 Diagram of the unified framework for subspace based face recognition.

also beneficial to compact $\tilde{I}$ onto a small number of features. We now analyze the behavior of the three subspaces for PCA, LDA and Bayes in order to discover how they suppress the $\tilde{T}$ and $\tilde{N}$ components, and compact $\tilde{I}$ component in their respective subspaces.

3.1 PCA eigenspace

Eigenfaces are computed from the ensemble covariance matrix C. Equation (2-1) shows that C is derived from all training face images subtracting of the mean face. We will show that C also can be computed from the face difference set, $\{(\vec{x}_i - \vec{x}_j)\}$, containing all the differences between any pair of face images in the training set.

Theorem 1. The eigenspace of PCA characterizes the difference between any two face images, which may belong to the same individual or different individuals.

Proof. In order to prove Theorem 1, we only need to show that the covariance matrix C for the image set $\{\vec{x}_1,\ldots\vec{x}_M\}$ can also be computed as

$$C = \frac{1}{2M}\sum_{i=1}^{M}\sum_{j=1}^{M}(\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T .$$

Since C is the ensemble covariance matrix for the training face images, from Eq. (2-1), we have

$$C = \sum_{i=1}^{M}(\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m})^T .$$

Replace $\bar{m}$ with Eq. (2-2),

$$C = \sum_{i=1}^{M}\left(\bar{x}_i - \frac{\bar{x}_1 + \ldots + \bar{x}_M}{M}\right)\left(\bar{x}_i - \frac{\bar{x}_1 + \ldots + \bar{x}_M}{M}\right)^T$$

$$= \frac{1}{M^2}\sum_{i=1}^{M}\left\{\left[(\bar{x}_i - \bar{x}_1) + \ldots + (\bar{x}_i - \bar{x}_M)\right]\cdot\left[(\bar{x}_i - \bar{x}_1) + \ldots + (\bar{x}_i - \bar{x}_M)\right]^T\right\}$$

$$= \frac{1}{M^2}\sum_{i=1}^{M}\left[\sum_{j=1}^{M}\sum_{k=1}^{M}(\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_k)^T\right] \tag{3-9}$$

Rewrite C using different subscripts (exchange i and j),

$$C = \frac{1}{M^2}\sum_{j=1}^{M}\left[\sum_{i=1}^{M}\sum_{k=1}^{M}(\bar{x}_j - \bar{x}_i)(\bar{x}_j - \bar{x}_k)^T\right].$$

Change the order of summation,

$$C = \frac{1}{M^2}\sum_{i=1}^{M}\left[\sum_{j=1}^{M}\sum_{k=1}^{M}(\bar{x}_j - \bar{x}_i)(\bar{x}_j - \bar{x}_k)^T\right]. \tag{3-10}$$

Average (3-9) and (3-10),

$$C = \frac{1}{2}\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M}(\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_j)^T$$

$$= \frac{1}{2M}\sum_{i=1}^{M}\sum_{j=1}^{M}(\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_j)^T. \tag{3-11}$$

Removing the scale $\frac{1}{2M}$ will not affect the eigenvectors of C, thus

$$C = \sum_{i=1}^{M}\sum_{j=1}^{M}(\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_j)^T \tag{3-12}$$

Therefore, the eigenvectors for the training set $\{\bar{x}_i\}$ can also be computed as the eigenvectors for the set of face difference $\{(\bar{x}_i - \bar{x}_j)\}$. This shows that the PCA eigenspace characterizes the variations between any two face images in the training data set.

3.2 Intrapersonal and extrapersonal subspaces

In the Bayesian algorithm, intrapersonal subspace and extrapersonal subspace are used to characterize the two kinds of variation Ω_I and Ω_E. The eigenvectors of intrapersonal subspace are computed from the image difference set $\{(\bar{x}_i - \bar{x}_j) \mid \ell(\bar{x}_i) = \ell(\bar{x}_j)\}$, for which the covariance matrix is

$$C_I = \sum_{\ell(\bar{x}_i)=\ell(\bar{x}_j)} (\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_j)^T . \tag{3-13}$$

The eigenvectors of extrapersonal subspace are derived from the difference set $\{(\bar{x}_i - \bar{x}_j) | \ell(\bar{x}_i) \neq \ell(\bar{x}_j)\}$, with covariance matrix

$$C_E = \sum_{\ell(\bar{x}_i)\neq\ell(\bar{x}_j)} (\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_j)^T . \tag{3-14}$$

Comparing C_I and C_E with C, we derive the following theorem,

Theorem 2. The intrapersonal subspace and extrapersonal subspace are the two components of the PCA eigenspace, and the extrapersonal eigenfaces are similar to the PCA eigenfaces.

Proof. From Eq. (3-12), we have

$$\begin{aligned} C &= \sum_{i=1}^{M}\sum_{j=1}^{M} (\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_j)^T . \\ &= \sum_{\ell(\bar{x}_i)=\ell(\bar{x}_j)} (\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_j)^T + \sum_{\ell(\bar{x}_i)\neq\ell(\bar{x}_j)} (\bar{x}_i - \bar{x}_j)(\bar{x}_i - \bar{x}_j)^T \\ &= C_I + C_E \end{aligned} \tag{3-15}$$

C is composed of C_I and C_E. Therefore the intrapersonal subspace and extrapersonal subspace are simply the two components of the standard eigenspace. Since the sample number for C_E is far greater than the sample number of C_I, the energy of C_E usually dominates the computation of C. Therefore, the extrapersonal eigenfaces are similar to the standard eigenfaces.

In Ω_E, $\tilde{T}$ and $\tilde{I}$ are coupled together. Therefore as discussed later, the extrapersonal subspace, which is similar to the standard eigenspace, cannot contribute much to separating $\tilde{T}$ and $\tilde{I}$. In fact, the improvement of the Bayesian algorithm over the PCA method benefits mostly from the intrapersonal subspace. The ML measure using the intrapersonal subspace alone is almost as effective as the MAP measure using the two subspaces [10]. So we will focus on intrapersonal subspace and the ML measure for the Bayesian algorithm in the later discussion.

3.3 LDA subspace

The subspace for LDA is derived from the within-class scatter matrix and the between-class scatter matrix. Similar to the analysis of the PCA and Bayesian

approaches in the previous sections, we can also study the LDA subspace using image difference subspace.

Theorem 3. The within-class scatter matrix is identical to C_I, the covariance matrix of the intrapersonal subspace, which characterizes the distribution of face variation for the same individuals. Using the mean face image to describe each individual class, the between-class scatter matrix characterize the variation between any two mean face images.

Proof. For simplicity, we assume that each class has the same sample number n. Similar to the proof of Theorem 1, we have,

$$\begin{aligned} S_w &= \sum_{i=1}^{L} \sum_{\vec{x}_k \in X_i} (\vec{x}_k - \vec{m}_i)(\vec{x}_k - \vec{m}_i)^T \\ &= \frac{1}{2n} \sum_{i=1}^{L} \sum_{\vec{x}_{k_1}, \vec{x}_{k_2} \in X_i} (\vec{x}_{k_1} - \vec{x}_{k_2})(\vec{x}_{k_1} - \vec{x}_{k_2})^T \\ &= \frac{1}{2n} \sum_{\ell(\vec{x}_i)=\ell(\vec{x}_j)} (\vec{x}_i - \vec{x}_j)(\vec{x}_i - \vec{x}_j)^T . \end{aligned} \tag{3-16}$$

Therefore,

$$S_w = C_I .$$

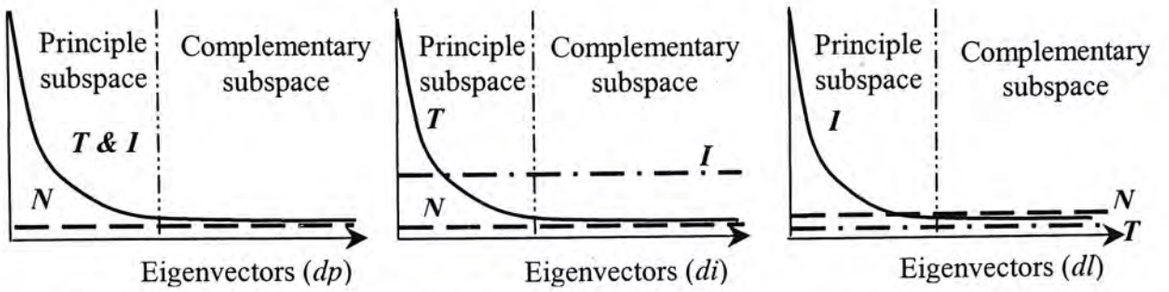
$$\begin{aligned} S_b &= \sum_{i=1}^{L} n(\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T \\ &= \frac{n}{2L} \sum_{i=1}^{L} \sum_{j=1}^{L} (\vec{m}_i - \vec{m}_j)(\vec{m}_i - \vec{m}_j)^T . \end{aligned} \tag{3-17}$$

This shows that S_b is the covariance matrix of the face difference set $\{(\vec{m}_i - \vec{m}_j)\}$.

3.4 Comparison of the three subspaces

PCA and LDA are initially developed considering class variation. According to the three theorems, they also can be illustrated in the view of face difference as Bayes. We now can compare these subspaces on how to process the face difference model. As mentioned earlier, a good subspace for recognition should be able to separate discriminating information $\tilde{I}$ from the deteriorating factors $\tilde{T}$ and $\tilde{N}$, and compact $\tilde{I}$ into a small number of features.

We first look at the PCA subspace as shown in Fig. 3-2 (a). Since PCA subspace characterizes difference between any two face images, it concentrates



(a) PCA subspace (b) Intrapersonal subspace (c) LDA subspace

Figure 3-2 Energy distribution of three components $\tilde{I}$, $\tilde{T}$, and $\tilde{N}$ on eigenvectors in three subspaces.

both $\tilde{T}$ and $\tilde{I}$ as structural signals on a small number of principal eigenvectors. By selecting the principal components, most noise encoded on the large number of trailing eigenvectors is removed from $\tilde{T}$ and $\tilde{I}$. PCA also compacts the feature space, since many dimensions where face images have almost zero projections have been removed. However, because of the continuing presence of $\tilde{T}$, the PCA subspace is not ideal for face recognition.

For the Bayesian algorithm, the intrapersonal subspace plays a critical role, while the extrapersonal subspace cannot contribute much to separating $\tilde{T}$ and $\tilde{I}$ since it is similar to the PCA subspace containing both $\tilde{T}$ and $\tilde{I}$ as structural signal. Since intrapersonal variation only contains $\tilde{T}$ and $\tilde{N}$, PCA on the intrapersonal variation arranges the eigenvectors according to the energy distribution of $\tilde{T}$, as shown in Fig. 3-2 (b). When we project a face difference Δ (either intrapersonal or extrapersonal) onto the intrapersonal subspace, most energy of the $\tilde{T}$ component will concentrate on the first few largest eigenvectors, while the $\tilde{I}$ and $\tilde{N}$ components are randomly distributed over all of the eigenvectors. This is because $\tilde{I}$ and $\tilde{N}$ are somewhat independent of $\tilde{T}$, which forms the principal vectors of the intrapersonal subspace. In Eq. (3-3) and (2-21), the Mahalanobis distance in F weights the feature vectors by the inverse of eigenvalues. This effectively reduces the $\tilde{T}$ component since the principal components with large eigenvalues are significantly diminished. $\varepsilon^2(\Delta)$ is also a distinctive feature for recognition, since it throws away most of the component $\tilde{T}$ on the largest eigenvectors, while keeps the majority of $\tilde{I}$, in the complementary subspace $\overline{F}$.

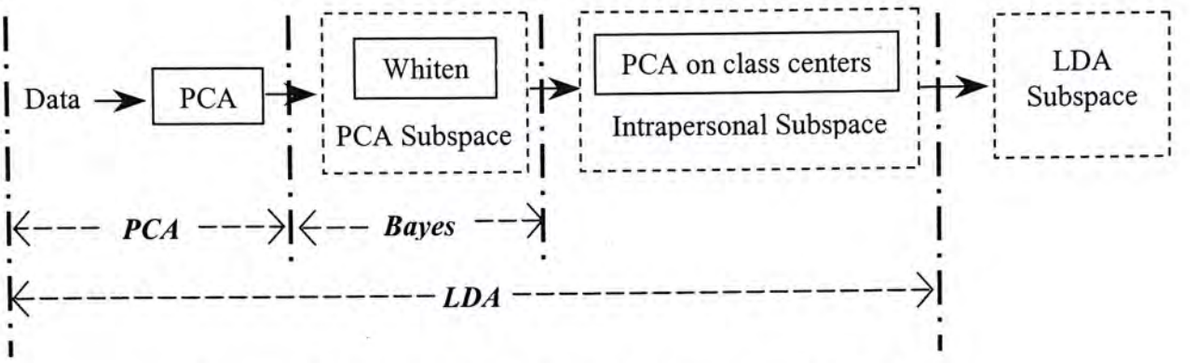


Figure 3-3 Relationship of the PCA, Bayes, and LDA subspaces.

The Bayesian algorithm successfully separates $\tilde{T}$ from $\tilde{I}$. However, $\tilde{I}$ and $\tilde{N}$ are still coupled on the small eigenvectors. Even though $\tilde{N}$ is usually of small energy, when it is normalized by the small eigenvalues in Eq. (2-21) and (3-3), the effect of $\tilde{N}$ could be significantly enlarged in the probabilistic measure. Another drawback for the Bayesian algorithm is that the intrinsic difference is not compacted, spreading over F and $\bar{F}$. It leads to high computation cost. In Bayes, DFFS requires computing the reconstruction error at every match, and its computation cost is equivalent to the correlation of two high dimensional vectors. An efficient computation using only DIFS is proposed in [10]. But when the feature number is small, its performance is poor, since the main component in F is $\tilde{T}$.

Finally, we look at the LDA subspace. The LDA procedure can be divided into three steps. First, PCA is generally used to reduce the data dimensionality. As discussed earlier, noise N is significantly reduced in this step. In the second step, to whiten the within-class scatter matrix we first compute its eigenvector matrix Φ and eigenvalue matrix Θ. From Theorem 3, we know that Φ spans the intrapersonal subspace, therefore Θ essentially represents the energy distribution of $\tilde{T}$. The whitening process projects data onto intrapersonal subspace Φ and normalizes them by $\Theta^{-1/2}$. We can see that this process is essentially the same as the Bayesian analysis. It reduces $\tilde{T}$ in the same manner.

In the third step of LDA, the PCA is again applied on the whitened class centers. In the process of averaging images in each class to compute the class centers, the noise $\tilde{N}$ is further reduced in this step. This is useful since $\tilde{N}$ may have been enlarged to a certain degree in the second step whitening process. Since both $\tilde{T}$ and $\tilde{N}$ have been reduced up to this point, the main energy in the class centers is

Table 3-1 Behavior of subspace on characterizing the face image difference.

Algorithm	Subspace	Decompose Face Image Difference	
		Principal space	Complementary space
PCA	Eigenspace	$\tilde{T}+\tilde{I}$	$\tilde{N}$
Bayes	Intrapersonal subspace	$\tilde{T}$	$\tilde{I}+\tilde{N}$
	Extrapersonal subspace	$\tilde{T}+\tilde{I}$	$\tilde{N}$
LDA	LDA subspace	$\tilde{I}$	$\tilde{T}+\tilde{N}$

the intrinsic difference $\tilde{I}$. However, as shown in Fig. 3-2 (b), $\tilde{I}$ is obtained by discarding principal component $\tilde{T}$ in the intrapersonal subspace, so $\tilde{I}$ spreads over the entire eigenvector axis after the whitening. The PCA on the class centers therefore serves two purposes. First, it can further reduce the noise as PCA usually does. Second, it compacts the energy of $\tilde{I}$ onto a small number of principal components, as shown in Fig. 3-2 (c).

Finally, the subspace analysis results of the three methods on the image difference model are summarized in Table 3-1. Instead of a simple combination of the three methods, the main contribution of our subspace analysis is to study the unique contribution of each subspace to the processing of face difference model. The degree of control over the $\tilde{I}$, $\tilde{T}$ and $\tilde{N}$ components in the face image difference depends on the dimensionality of the three subspaces, the PCA subspace (*dp*), intrapersonal subspace (*di*), and LDA subspace (*dl*).

3.5 L-ary versus binary classification

Under this framework, we find that face recognition can be treated as a binary classification problem for intrapersonal and extrapersonal variations, instead of a *L*-ary classification problem. It is the fundamental difference from other pattern recognition problems, and critical for the success of subspace based methods. For normal pattern recognition, when there are enough samples for each class, a Bayesian classifier can be used based on the covariance matrices estimated for each individual class,

$$L(\bar{x},\bar{m}_i)=\frac{1}{2}(\bar{x}-\bar{m}_i)^T\Gamma_i^{-1}(\bar{x}-\bar{m}_i)+\ln(|\Gamma_i|) \tag{3-18}$$

where Γ_i is the covariance matrix for each pattern class.

However, for face recognition, there are usually too few samples for each class to correctly estimate the class covariance matrix. Therefore, it is difficult to use a

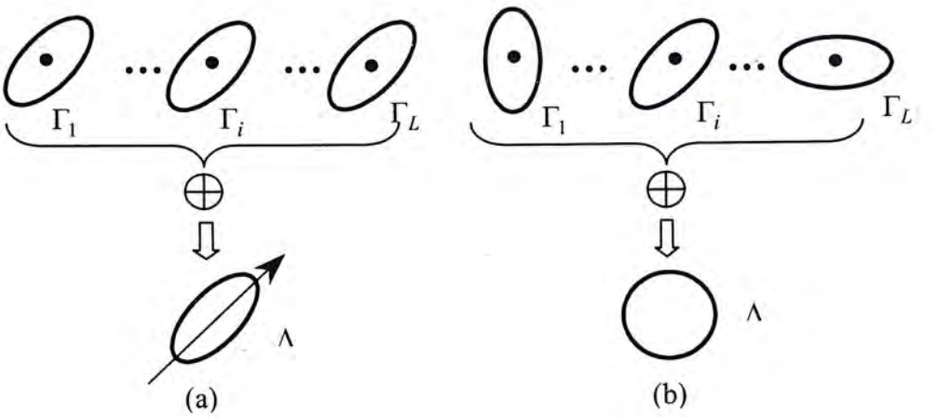


Figure 3-4 Use average intrapersonal variation distribution to approximate that for each individual class.

When the class scatter matrices are consistent like (a), the average intrapersonal variation distribution can be used to approximate to that for each individual class. Otherwise like (b), it is not appropriate to replace Γ_i with Λ.

L-ary Bayesian classifier directly. Fortunately, human faces are not only similar in structure, but also in facial variations. We share similar facial expressions. This means people tend to have similar intrapersonal variation. Thus we can pool a large number of face classes together to estimate an average covariance matrix Λ to reflect the intrapersonal variation. As shown in Fig. 3-4 (a), when all of the individuals have similar scatter matrix, spanning in the same direction, Λ is a good approximation of Γ_i. This is generally not the case for many pattern recognition problems where sample distribution for each class is different from each other as shown in Fig. 3-4 (b).

3.6 Unified subspace analysis

From this framework, it is found that that PCA and Bayes can be viewed as intermediate steps of LDA. However, conventional LDA does not attain the best performance possible without improving each individual step. The subspace dimension of each subspace method can affect the recognition performance. It is a trade-off on how much noise and transformation difference are included, and how much intrinsic difference is included. It also implies that the intrapersonal variation of one face can be estimated from the samples of other faces. Based on these considerations, we propose a unified subspace analysis method for face recognition as follows:

(1) Project face vectors to PCA subspace and adjust the PCA dimension (*dp*) to reduce most noise.

(2) Apply Bayesian analysis in the reduced PCA subspace and adjust the dimension (*di*) of intrapersonal subspace. Since human faces share similar intrapersonal variation, the transformation $\tilde{T}$ for a testing individual can be estimated from faces of others. Therefore, different from the standard subspace methods, our intrapersonal subspace is computed from an enlarged intrapersonal difference set that contain individuals both inside and outside of the gallery, so that the intrapersonal subspace is robust to all the transformations in the test set. PCA subspace is also computed from this enlarged training set.

(3) For the L individuals in the gallery, compute their training data class centers. Project all the class centers onto the intrapersonal subspace, and then normalize the projections by intrapersonal eigenvalues to compute the whitened feature vectors.

(4) Apply PCA on the whitened feature vector centers to compute a discriminant feature vector of dimension dl.

This algorithm has two major improvements over traditional subspace methods. First, it provides a new parameter space to improve recognition performance. The method controls the $\tilde{I}$, $\tilde{T}$ and $\tilde{N}$ components in the image difference by adjusting the dimensionality of the three subspaces. The interaction of the three parameters greatly affects the system performance. Using each of the three subspace dimensions as a parameter axis, the algorithm provides a three-dimensional parameter space, as shown in Fig. 3-5.

The original PCA, LDA, and Bayes methods only occupy some local lines or areas in the 3D parameter space. PCA changes parameters in the *dp* direction on line *AD*. DIFS and DFFS of the Bayesian algorithm change on the line *DEF* in the *di* direction. Fisher Face [58] corresponds to point *B* (*dp=di=M-L*, *dl=L-1*) in the graph. All these methods change parameters only in the local regions. However, for our new algorithm, optimal parameters may be searched in the full 3D space. We can clearly see the advantage of this in the experiments.

The second improvement of the algorithm is the adoption of different training data at different steps of the training process according to the special requirement of the step. In traditional method, the same training data is used throughout the algorithm. The conflict requirements of each step limit the optimization ability of the algorithm. For example, in the LDA method, S_w and S_B come from the same

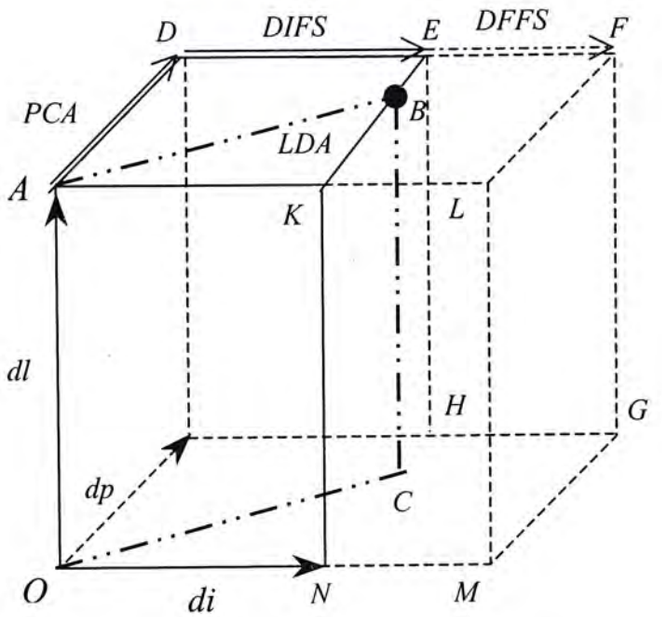


Figure 3-5 3D parameter space.

dp, *di*, and *dl* are the dimensionalities of PCA subspace, intrapersonal subspace, and LDA subspace.

training data. Normally, only the individuals in the gallery are selected for training. The samples for each class may be too few to estimate the transformation difference $\tilde{T}$ to be appeared in testing, since sometimes there is only one sample for each individual in the gallery. When there are not enough training samples, the intrapersonal eigenvectors with very small eigenvalues are sensitive the slight change on training set, and the LDA classifier is unstable. Including the training samples outside gallery can improve the PCA subspace and the intrapersonal subspace. However, if we add to the training set with many more individuals who are not in the gallery, the between-class scatter matrix S_B maybe too distracted to extract optimal features targeting the discrimination of the individuals in the gallery.

In order to accomodate this conflicting requirements, we use different training set for different steps. For the PCA and intrapersonal subspace estimation (step-1,2) we use an enlarged intrapersonal difference set that contain individuals both inside and outside of the gallery to effectively estimate $\tilde{T}$ and $\tilde{N}$. Then for the discriminant analysis step (step-3,4), we only use the class centers of the individuals who are in the gallery, so that the features extracted are specially tuned for the individuals in the gallery.

3.7 Discussion

Starting from a the face difference model that decomposes a face difference into three major components, intrinsic difference $\tilde{I}$, transformation difference $\tilde{T}$, and noise $\tilde{N}$, we unify the three major subspace face recognition methods, PCA, Bayes, and LDA under the same framework. Using this framework we discover how each of the three methods contributes to the extraction of discriminating information $\tilde{I}$ in the face difference. This eventually leads to the construction of a 3D parameter space that use the three subspace dimensions as axes. Searching through this parameter space, we achieve better recognition performance than the standard subspace methods.

This framework provides a better understanding on how to select proper training set for face recognition. We find that when computing the intrapersonal subspace, the training set should contain the transformations that may appear in the test set of the application. Especially, since the intrapersonal variation for one individual can be estimated from that of others, to compute the intrapersonal eigenspace for the Bayes and LDA methods we can add the samples of individuals not in the gallery into the training set. On the other hand, in the third step of the LDA algorithm, the between-class matrix is used to extract discriminating difference among different individuals. It is better that the samples computing the between-class matrix come from the same individuals as the ones in the gallery.

Under this framework, we can discover many parameter regions unexplored by previous research. For example, as we discussed earlier, LDA is performed based on the Bayesian analysis in the intrapersonal eigenspace. Since the intrapersonal complementary subspace also contains some distinctive features for recognition, we can easily extend the standard LDA to the complementary subspace. This effectively corresponds to the cube EFGH-KLMN in the parameter space as shown in Fig. 3-5. Further more, we can perform LDA in both the intrapersonal eigenspace and the complementary subspace, and then combine the two parts of discriminative features together.

However, there are also several problems open to discuss. One problem is how to find the optimal parameters. Searching through the whole 3D parameter space may be time consuming. A possible strategy is suggested as the steps of our

experiments in Section 4. First, observe the *dp-di* accuracy surface to decide *dp* and *di*, and the choose *dl* according to the accuracy curve in LDA subspace. For further simplicity, the range of *dp* can be first narrowed down by observing the PCA (Mahalanobis) accuracy curve. When much noise is included in the PCA subspace, the PCA (Mahalanobis) accuracy will greatly drop. This stagery can lead to some good parameters, although they are not necessarily optimal.

In this framework, the intrapersonal variation is modeled as Gaussian distribution, and it is assumed that $\tilde{T}$ and $\tilde{I}$ can be separated by PCA on intrapersonal difference set. However, for significant $\tilde{T}$, such as large changes in pose, this assumption may break down. The face difference may not be modeled as linear composition of $\tilde{I}$, $\tilde{T}$, $\tilde{N}$. This is also a crucial problem to standard subspace methods and better understanding of this framework. To solve this problem, one way is to "normalize" the large lighting and pose changes using such approaches as 3D model before subspace analysis. Another way is to model the intrapersonal difference as more complex distribution, such as Gaussian Mixture Model (GMM). It may also be helpful to apply ICA, or kernel PCA to face image before further linear subspace analysis, since they can address the high order dependencies of different factors.

Another assumption for this framework is that human face share similar structural and intrapersonal variation. It is not suitable for general pattern recognition problems. So we should be careful to extend this framework to other applications. For example, it is usually not desirable to add extra samples not in the gallery into the training set.

Chapter 4

Experiments on Unified Subspace Analysis

In this section, we conduct experiments on two data sets from the FERET face database [56] and the AR face database [4] to evaluate the unified subspace analysis. In the preprocessing procedure, all the images are normalized for scaling, translation, and rotation, such that the eye centers are in fixed positions. A 27×41 mask template is used to remove the background and most of the hair. Histogram equalization is applied to the face images for photometric normalization. An example of the normalized face image is should in Figure 4-1. Before subspace analysis, the image vector is normalized to zero mean and unit variance.

4.1 Experiments on FERET database

In the first data set, we select 1195 persons from the FERET database, with two face images (FA/FB) for each person. Images of 495 persons are used for training, and the remaining 700 persons are used for testing. So there are totally 990 face images in the training set, 700 face images in the gallery and 700 face images for probe. This data set is selected so that the individuals for training and testing are separated, and there is a large class number with a small sample number for each class.

4.1.1 PCA Experiment

We use the Euclid and Mahalanobis distance measures for the PCA recognition. The recognition accuracy for different number of eigenvectors (*dp*) is shown in Fig. 4-2. The accuracy of direct correlation is 84.1%. We use direct correlation as a benchmark since it is essentially a direct use of image difference without subspace analysis. When *dp* is small, the PCA result with Euclid measure is worse than correlation. As *dp* increases, it steadily approaches the benchmark. There is no noticeable improvement when using the Mahalanobis measure. It reaches peak accuracy (84.3%) with around 150 eigenvectors, and then drops with further increase of dimensionality. For Mahalanobis measure, since each dimension needs

Figure 4-1 Example of normalized face image.

to be normalized by its corresponding eigenvalue, the high dimensional components with small eigenvalues are significantly magnified. Since these dimensions tend to contain more noise than structural signal, they will deteriorate the recognition results. This explains the drop of recognition accuracy when *dp* is increased. The overall results in Fig. 4-2 shows that PCA is no better than direct correlation in term of recognition accuracy. This confirms the analysis in Section 3.1 and Section 3.4. Even though PCA can effectively reduce subspace dimension and remove noise $\widetilde{N}$, it cannot decouple the intrinsic difference $\widetilde{I}$ and transformation difference $\widetilde{T}$ to improve recognition accuracy.

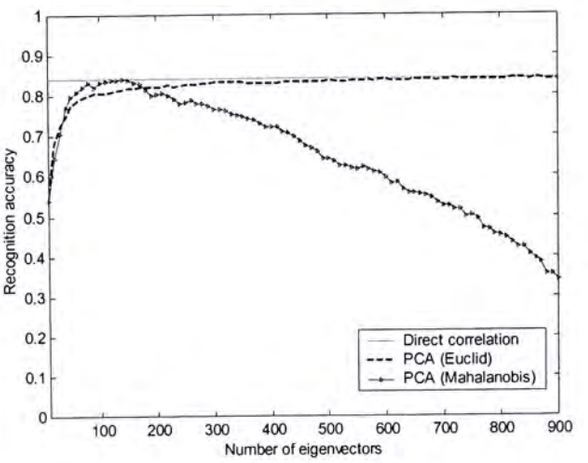


Figure 4-2 Recognition accuracy of PCA on the FERET database.

4.1.2 Bayesian experiment

Experimental results for the Bayesian algorithm are reported in Fig. 4-3. The Bayesian algorithm has achieved around 10% improvement over direct correlation. We notice that the Bayesian algorithm is stable for different intrapersonal eigenvector number *K*. The eigenvectors of the intrapersonal subspace are arranged by the energy of $\widetilde{T}$. When only a small number of eigenvectors are

selected, the principal space does not have enough information on $\tilde{I}$, so the accuracy of DIFS is low (below 60% for 20 eigenvectors). The lost information can be compensated from DFFS in the complementary space. So the accuracy of ML is high, around 93%, since it combines the two components together. However, even for small K, the computation cost of ML and DFFS is high, equivalent to correlation of two high dimensional vectors. When we use Euclid instead of Mahalanobis distance measure for DIFS, the recognition accuracy drops greatly, and becomes even worse than PCA. Since the main component in the intrapersonal eigenspace is $\tilde{T}$, without using the eigenvalues to reduce $\tilde{T}$, the Euclid distance has to compute face difference mainly based on $\tilde{T}$. On the contrary, the eigenvalue normalization of the Mahalanobis in PCA does not help to improve the PCA recognition accuracy as shown in Fig. 4-2. This is because in the PCA subspace, $\tilde{T}$ and $\tilde{I}$ are coupled together, thus the eigenvalues cannot reflect the energy distribution of $\tilde{T}$ alone in order to effectively reduce $\tilde{T}$.

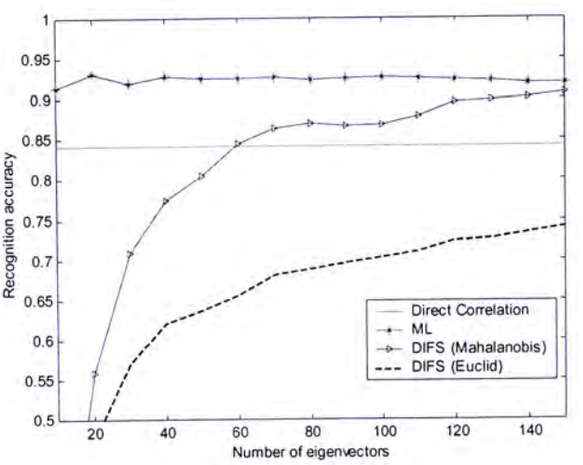


Figure 4-3 Recognition accuracy of the Bayesian algorithm on the FERET database.

4.1.3 Bayesian analysis in reduced PCA subspace

After comparing the PCA and Bayesian methods individually, we now use a set of experiments to investigate how these two subspace dimensions in our 3D parameter space may interact with each other. We first apply PCA on the raw face vector to reduce the dimensionality and remove the noise. Then the Bayesian analysis is implemented in the reduced PCA space. This corresponds to the *dp-di* plane in the 3D space in Fig. 3-5.

Results are reported in Table 4-1. The vertical direction is the dimension of the PCA subspace (*dp*) and the horizontal direction is the dimension of the intrapersonal subspace (*di*), refered as the intrapersonal eigenvector number K in Eq. (2-21). For better viewing of the results, the *dp-di* accuracy surface is also plotted in Fig. 4-4. There are two benchmark curves in the 3D space of Fig. 4-4. One is accuracy curve of the traditional PCA method as reported in the second column in Table 4-1. It is used to evaluate the improvement of the Bayesian analysis. The second curve is the DIFS curve of the standard Bayesian algorithm based on raw face vectors, equivalent to the DIFS (Mahalanobis) curve in Figure 4-3. It is reported in the bottom row of Table 4-1. We compare it with DIFS curves in different PCA subspaces. Since there are 990 face images and 495 classes in the training set, the rank of the within-class scatter matrix is bounded by 495. The maximum value for *di* is $\min\{d_p, 495\}$.

The shape of the *dp-di* accuracy surface clearly reflects the effect of noise $\tilde{N}$. When *dp* is small, there is little noise in the PCA subspace. So the recognition accuracy monotonically increases with *di* as more discriminating information $\tilde{I}$ is added, and finally reaches the highest point at the full dimensionality of the intrapersonal subspace. However, as *dp* increases, noise begins to appear in the PCA subspace and causes a change in the accuracy curve shape. The curve starts to decrease after reaching a peak point before *di* reaches the full dimensionality. The decrease in accuracy at the end of the curve is because noise distributed on the small eigenvectors is magnified by the inverse of the small eigenvalues.

This effect of noise is especially severe when both *dp* and *di* are around 495, i.e. the largest possible *di*. In this region, the accuracy becomes as low as 67%. Because of the large *dp*, noise has become a fairly significant problem. When *di* becomes the same size as *dp*, all the energy in the PCA subspace, including noise, are selected for the Bayesian analysis. Noise concentrated on the last few very small eigenvectors will be drastically magnified because of the very small eigenvalues. Therefore, we observe a low accuracy region around the area where both *dp* and *di* equal 495. Interestingly, for this training set, the parameters proposed in Fisher face [42] actually falls into this region. This shows the importance of the parameter selection for a given training set.

We plot the highest accuracy of each accuracy curve of different *dp* in Fig. 4-5. The maximum point with 96% accuracy could be found at (*dp*=150, *di*=150). In this PCA subspace, noise has been removed and all of the eigenvectors can be used for Bayesian recognition. A pre-step of PCA can improve the performance of Bayes. More experimental results can be found in our previous work [89].

Table 4-1 Recognition accuracy of Bayesian analysis in reduced PCA subspace (%).

	Euclid	*Dp*	DIFS								
			10	20	50	100	150	200	250	300	490
PCA	77.3	50	27.7	60.9	93.7	N/A	N/A	N/A	N/A	N/A	N/A
	80.7	100	27.1	58.1	85.4	95.4	N/A	N/A	N/A	N/A	N/A
	81.7	150	27.6	57.3	81.4	90.9	96.0	N/A	N/A	N/A	N/A
	82.1	200	27.6	58.0	81.3	89.3	92.3	95.3	N/A	N/A	N/A
	82.7	250	26.9	57.1	80.6	87.7	91.0	94.9	94.4	N/A	N/A
	83.1	300	27.1	56.7	80.6	87.9	93.7	93.7	94.4	93.0	N/A
	82.9	400	26.7	56.6	80.3	87.1	91.0	92.3	92.9	94.3	N/A
	83.6	500	26.6	56.3	80.4	87.1	90.7	91.6	92.7	93.1	67.0
	83.9	600	26.6	56.1	80.3	86.9	90.7	91.9	92.6	92.3	89.7
	84.0	700	26.7	56.0	80.3	86.9	90.7	92.0	92.6	93.1	91.1
	84.0	900	26.6	56.0	80.4	86.9	90.7	91.7	92.6	93.0	90.9
Bayes on raw data			26.7	55.9	80.4	86.9	90.7	91.9	93.0	93.0	90.6

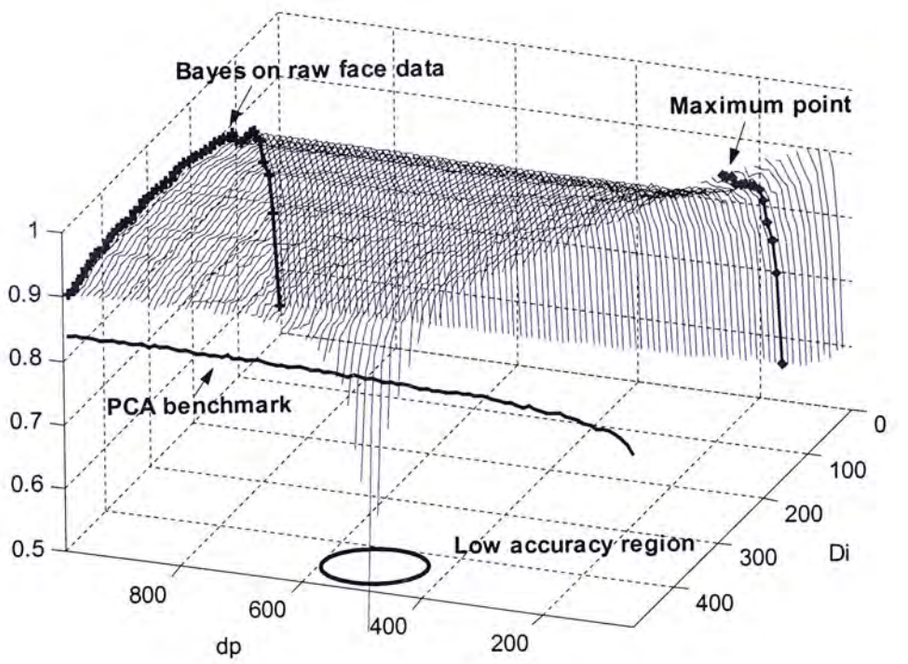


Figure 4-4 Accuracy surface for the Bayesian analysis in the PCA subspace.

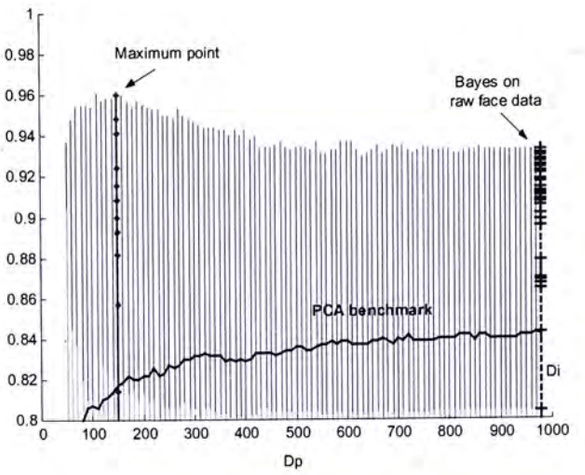


Figure 4-5 Highest accuracy of the Bayesian analysis in each PCA subspace.

4.1.4 Extract discriminant features from intrapersonal subspace

We now investigate the effect of the third dimension *dl* in the 3D parameter space. For ease of comparison, we choose four representative points on the *dp-di* surface, and report the accuracy along the dimensions of *dl* as shown in Fig. 4-6. The curves first increase to a maximum point and then drop with further increase of *dl*. For traditional LDA, the *dl* dimension is usually chosen as $L-1$, which corresponds to the last point of the curve with *di* = 495. The result is clearly much lower than the highest accuracy in Fig. 4-6. As discussed in Section 3, this dimension mainly serves to compact $\tilde{I}$ and remove more noise $\tilde{N}$, so the dimensionality should be reasonably small instead of being fixed by L. The best results on the plots are indeed better than using the first two dimensions only. Figure 4-7 compares the recognition accuracies using small feature number for each step of the framework. For Bayes, DIFS measure is used for comparison, since ML measure is in high computation cost even for small feature number. It clearly demonstrates the improvement on recognition efficiency.

As shown by these experiments, although we have not explored the entire 3D parameter space, better results are already found comparing to the standard subspace methods. A careful investigation of the entire parameter space should lead to further improvement.

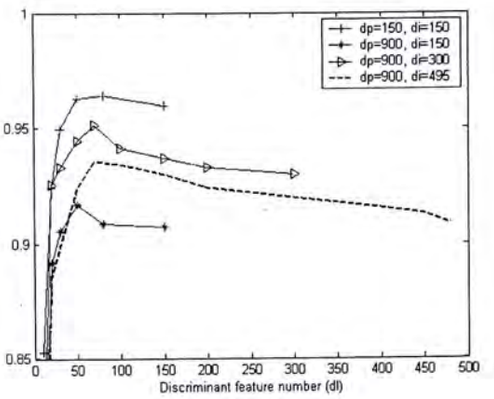


Figure 4-6 Accuracies using different number of discriminant features extracted from intrapersonal subspace.

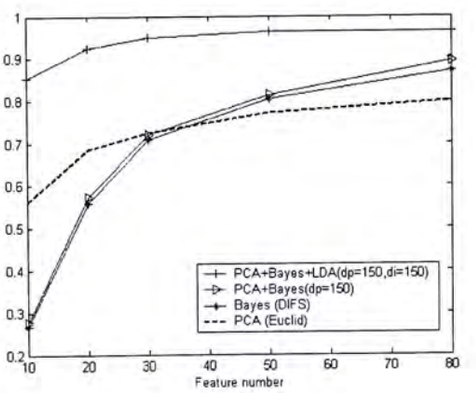


Figure 4-7 Recognition accuracies using small feature number for each step of the framework.

4.1.5 Subspace analysis using different training sets

As discussed earlier, different training sets can be used in the steps of framework. To better illustrate this improvement, a small data set is constructed from the FERET database. The data set contains 100 persons, and there are four face images taken in two different sessions for each person. Two face images are in gallery, and another two are for probe. Although the data set size is much smaller, recognizing face images of different sessions usually is much more difficult than recognition of FA/FB set with only expression changes.

As shown in Figure 4-8, we study the subspace analysis performance in three different ways of using training sets. First, in case (I), we use the 200 samples in the gallery as training set throughout the steps of subspace analysis. It is also the conventional way for LDA. In the experiment, the optimal parameter on the *dp-di* accuracy surface is found at (*dp*=130, *di*=70). The accuracy curve with difference

discriminative feature number is plotted in Figure 4-8. The performance is poor, because training set is too small. To reinforce the training set, we add another 1204 samples of 400 persons outside the gallery to the training set. The PCA subspace and the intrapersonal subspace are computed from the 1404 samples of 500 persons. By Bayesian analysis in the PCA subspace, we choose *dp=100*, *di=100*. The only difference between case (II) and case (III) is the between-class scatter matrix (S_b). In case (II), S_b is computed from class centers of all the 500 persons, while in case (III), S_b is just computed from the class centers of 100 people in gallery. The performance of both (II) and (III) is much better than (I), because the extra training samples outside the gallery effectively improve the PCA subspace and intrapersonal subspace. Because human faces share similar intrapersonal variation, the transformation difference in the probe set can be more accurately estimated using the larger training set. However, the performance of (III) is even better than (II). The discriminative features are more compacted on small number of features. In case (II), the S_b is computed from 500 people. The additional data serves mainly as distraction for the extraction of optimal features discriminating the 100 people in the gallery. In case (III), only the centers of classes in the gallery are used to compute S_b, the derived features are specially tuned for the people to be recognized. This improvement is notable when the gallery size is relatively small. It is much easier to find the discriminative features recognizing 100 people than 500 people.

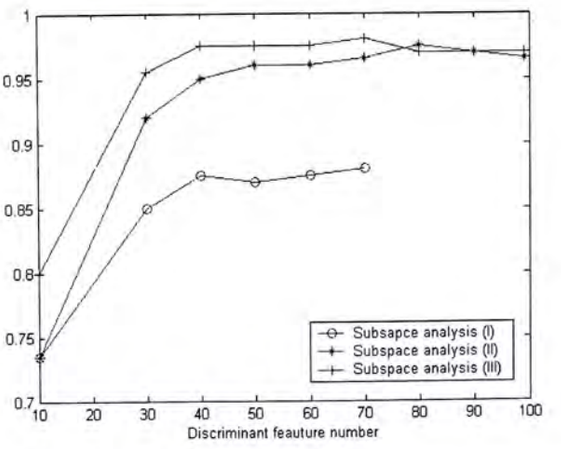


Figure 4-8 Subspace analysis for different training sets.

(I): All the three steps use the 200 samples of 100 people in the gallery for training.(II): All the three steps use the 1404 samples of 500 people including 400 people outside the gallery for training. (III) PCA subspace and intrapersonal subspace are computed from the 1404 samples, and the LDA subspace (between-class scatter matrix) are computed from only the 200 samples in the gallery.

4.2 Experiments on the AR face database

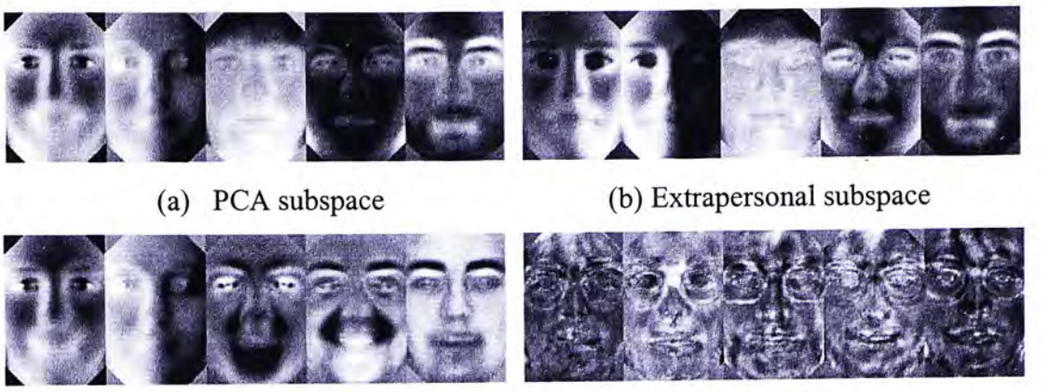
In order to study the properties of the framework under different data conditions, we conduct a second set of experiments using a data set from the AR-face database. Ninety people are selected from the AR-face database. For each individual, 14 face images taken in two sessions are selected. For each session, there are 7 face images under 7 different transformations as listed in Table 4-2. Face images of the seven transformations for a sample individual are shown in Fig. 4-9. In this experiment, face images in the first session are used for training and gallery, and face images in the second session are used as probe set. Different from experiment on FERET database, in this experiment the individuals for training and test are the same. The training set has a smaller class number and a larger sample number under various transformations for each class. So we can evaluate the recognition performance under different transformations.

The first five eigenfaces for each subspace are shown in Fig. 4-10. Since the training face images are all from the same session, the main transformations come from variation of expression and lighting. In the first five eigenfaces for PCA, the first, second, and fourth eigenfaces characterize the intrapersonal lighting variation; the third and fifth eigenfaces characterize the hairstyle and moustache variation, belonging to the intrinsic extrapersonal difference of this training set. The first five eigenfaces for extrapersonal eigenspace in Fig. 4-10 (b) are similar to those of the standard eigenspaces as pointed out in Theorem 2 earlier. The first five eigenfaces for intrapersonal eigenspace only characterizes intrapersonal expression and lighting transformations of the same individuals. The vectors for LDA on the other hand are not affected much by lighting and expression changes.

Table 4-2 Seven transformations for each individual class in each session from AR database.

Neutral	Expression			Lighting		
	Smile	Frown	Cry	Left	Right	Front
1	2	3	4	5	6	7

Figure 4.9 Samples for the seven transformations in AR database.



(a) PCA subspace (b) Extrapersonal subspace

(c) Intrapersonal subspace (d) LDA subspace

Figure 4-10 First five eigenfaces for different subspaces.

4.2.1 Experiments on PCA, LDA and Bayes

The 630 face images in the first session are used as training set to compute the standard PCA eigenspace, intrapersonal and extrapersonal subspaces for the Bayesian algorithm, and the LDA subspace. We report the best recognition accuracy of the three algorithms in Table 4-3. The PCA method uses the Euclid distance and the Bayesian algorithm uses the ML similarity measure for recognition. The 630 face images in the first session are used as the gallery and the 630 face images in the second session are used as the probe set.

Unlike experiment in Section 4.1, the Bayesian algorithm is not much better than the PCA method in this experiment. The transformation in the training set is mainly caused by lighting and expression. For every face image in the test set, there is a corresponding face image under the same transformation in the reference set. Using nearest neighbor classifier, the difference caused by lighting and expression is largely canceled in Δ. The main transformation factor in Δ affecting the recognition is caused by different sessions, e.g. the change of hairstyle, but not lighting and expression varying in the same session. The intrapersonal subspace cannot characterize the kind of transformation factor $\tilde{T}$ that needs to be overcome, so recognition cannot be much improved by the Bayesian algorithm. We will illustrate this point further in the later experiments.

The LDA method gives the best performance. The main contribution comes from the last step of LDA. Because there are only 90 individuals in both the training set and the gallery, it is easier to seek their difference by applying PCA to class centers. As reported in Table 4-4, we apply Bayesian analysis in the reduced PCA space, with 300 for D_p and D_i. If all of the 300 dimensions of the whitened intrapersonal subspace are used for recognition, the accuracy is only 80.8%. If

LDA selects 89 most distinctive features from the whitened intrapersonal space, then the accuracy is improved to 86%.

Table 4-3 Recognition result of PCA, LDA, and Bayes (ML) on AR databse.

	PCA	LDA	Bayes (ML)
Accuracy	81.6%	86.0%	82.9%

Table 4-4 Recognition accuracies of LDA on AR database using different feature number.

D_p	D_I	10	20	30	50	70	80	89	All
300	300	0.551	0.713	0.748	0.787	0.827	0.851	0.860	0.808

4.2.2 Evaluate the Bayesian algorithm for different transformation

Experiment in Section 4.2.1 implies that the transformation in training set should be consistent with that for test. We will evaluate the performance of the Bayesian algorithm under different transformations in this part. Conclusion is also suitable to LDA because of their relationship. We use the 90 neutral face images in the first session as gallery and 630 face images in the second session as probe set. Since the difference caused by lighting and expression cannot be canceled by the face difference, the recognition task is much more difficult than that in Section 4.2.1. The following experiment will illustrate how the Bayesian algorithm overcomes the expression and lighting transformations. The 630 face images for probe are divided into three groups. As described in Table 4-5, Probe set (I) includes Transformation 1, neutral face images. Probe (II) includes Transformation 2, 3, and 4, which is used to test the performance under expression variation. Probe set (III) includes Transformation 5, 6, and 7, used to test the performance under lighting variation. Selecting different face images in session one, we also design three training sets, which produce different intrapersonal subspace. Training set (I) includes Transformation 1, 2, 3, and 4. The subspace derived from this set is presumed to characterize the expression variation. Training set (II) includes Transformation 1, 5, 6, 7. The subspace derived from this set is expected to characterize the lighting variation. Training set (III) includes all of the seven transformations, so it is expected to characterize both expression and lighting variation.

The recognition accuracies on the three probe sets using direct correlation and the Bayesian algorithm based on the three different training sets is reported in Table 4-6 and Fig. 4-11. For all these algorithms, the accuracies on probe (I), neutral face images, change little. For direct correlation, the recognition accuracy drops greatly on probe (II) and (III) because of the difference between probe image and image in gallery caused by expression and lighting variation. The Bayesian algorithm can reduce the effect of the transformation to some extent, but different training sets lead to different performances. For training set (I), containing face images of different expressions, the accuracy on probe set (II) having expression variation has been greatly improved, but the accuracy on probe set (III) with lighting variation has no improvement to direct correlation. Similarly, Bayesian algorithm based on training set (II), improves the performance on probe (III), but is not effective on probe set (II). The face images in training set (III) contains both expression and lighting transformation, so the accuracy on both probe set (II) and (III) has been improved. So in order to improve the robustness of the Bayesian algorithm, training set must contains the kind of transformation that may appear in the test.

Table 4-5 The testing and training sets for different transformations.

		Transformation	Image Number
Training set	(I)	1,2,3,4	360
	(II)	1,5,6,7	360
	(III)	1,2,3,4,5,6,7	630
Testing set	(I)	1	90
	(II)	2,3,4	270
	(III)	5,6,7	270

Table 4-6 Recognition results of direct correlation and Bayesian algorithm for different transformations.

		Test (I)	Test (II)	Test (III)
Direct Correlation		85.6%	59.3%	28.5%
Bayes (ML)	Training (I)	88.9%	84.1%	28.9%
	Training (II)	91.1%	61.4%	87.4%
	Training (III)	88.7%	81.1%	75.6%

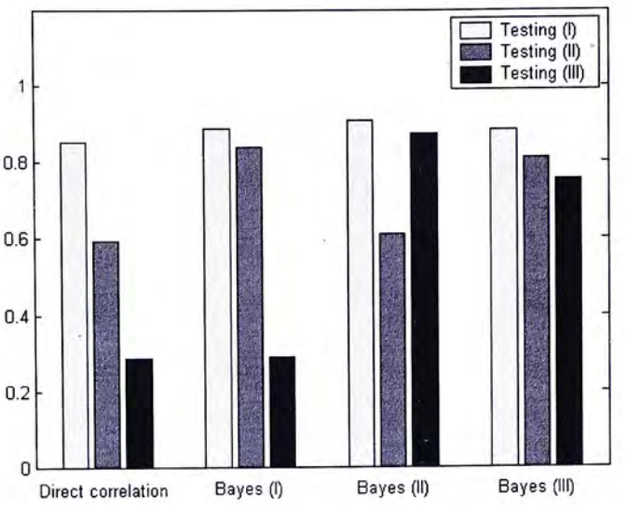


Figure 4-11 Accuracies of direction correlation and Bayesian algorithm for different transformations.

Chapter 5

Discriminant Analysis in Dual Subspaces

When dealing with the high dimensional face data, LDA often suffers from the small sample size problem. When there are not enough training samples, the within-class scatter matrix S_w may become singular, and it is difficult to compute the LDA vectors. It also may lead to the overfitting problem. LDA vectors are tuned to the training set with the existence of noise, since S_w is not well estimated. Our framework shows that the high dimensional image space can be decomposed into intrapersonal principal space, spanned by the eigenvectors of S_w, and its complementary subspace, also called the null space of S_w. One way to avoid matrix singularity is to first remove the null space of S_w. In a two-stage PCA+LDA [58][84], the data dimensionality was first reduced by PCA, and LDA was performed in the reduced PCA subspace, in which S_w is non-singular. In a enhanced LDA model proposed by Liu et. al. [16], a small set of eigenvectors of S_w was chosen, and the zero and trivial eigenvalues were excluded to avoid overfitting for noise. However, Chen et. al. [38] suggested that the null space spanned by the eigenvectors of S_w with zero eigenvalues contains the most discriminant information. A LDA method in the null space of within-class scatter matrix was proposed. It chooses the projection vectors maximizing the between-class scatter matrix with the constraint that S_w is zero. But this approach discards the discriminative information outside of the null space of S_w. Yu. et. al. [28] proposed a direct LDA algorithm, and claimed that it took advantage of all the information both within and outside of the null space of S_w. It first removes the null space of the between-class scatter matrix, and assumes that no discriminative information exists in this space. In this thesis, we find that the optimal discriminant vectors do not necessarily lie in the subspace spanned by the class centers. Considering all the above LDA approaches, they all lost some discriminative information in the high dimensional data space.

Our framework has proved that the Bayesian algorithm can be viewed as the intermediate step of LDA. Different from LDA, which extracts discriminative features in only one subspace, the Bayesian algorithm makes use of the information in two subspaces, but it does not further extract distinctive features to further separate the class centers. Moreover, in the Bayesian algorithm, computing the component in the intrapersonal complementary subspace is expensive, since it requires computing the reconstruction error. Starting from this framework, we propose a novel face recognition approach, which applies discriminant analysis in dual intrapersonal subspaces, and combines the two parts of discriminative features under a probabilistic model. It integrates the advantages of the Bayesian and LDA algorithms. This novel approach is also much more efficient than Bayes, since the matching in both of the two intrapersonal subspaces is based on low dimensional features. Experiments on the FERET database and the XM2VTS database [37] clearly demonstrate the superiority of this new method.

5.1 Review of LDA in the null space of S_w and direct LDA

In Section 2.2, we have reviewed the popular Fisherface [58] based on the PCA+LDA. In this section, before giving our novel approach, we will first review another two modified LDA approaches, LDA in the null space of S_w, and direct LDA.

5.1.1 LDA in the null space of S_w

The LDA approach in Section 2.2 is performed in the principal subspace of S_w, in which $W^T S_w W \neq 0$. However, the null space of S_w, in which $W^T S_w W = 0$, also contains much discriminative information, since it is possible to find some projection vectors W satisfying $W^T S_w W = 0$ and $W^T S_b W \neq 0$, and the Fisher criteria (2-10) definitely reaches its maximum value. A LDA in the null space of S_w was proposed by Chen et. al. [38]. It chooses the projection vectors maximizing S_b with the constraint that S_w is zero. First, the null space of S_w is computed as,

$$V^T S_w V = 0 \quad (V^T V = I) \tag{5-1}$$

The between-class scatter matrix S_b is projected to the null space of S_w,

$$\widetilde{S}_b = V^T S_b V . \tag{5-2}$$

Choose the eigenvectors U of $\widetilde{S}_b$ with the largest eigenvalues Λ,

$$U^T \widetilde{S}_b U = \Lambda . \tag{5-3}$$

The LDA transformation matrix is defined as $W = VU$.

This LDA approach utilizes the discriminative information in the null space of S_w. But unfortunately, as the rank of S_w increases, the null space of S_w becomes small, and much discriminative information outside it is discarded [28].

5.1.2 Direct LDA

Yu et. al. [28] proposed a direct LDA method, and it was claimed to take advantage of all the discriminative information within and outside of the null space of S_w. In this approach, S_b is first diagonalized, and the null space of S_b is removed,

$$Y^T S_b Y = D_b > 0 , \tag{5-4}$$

where Y are eigenvectors and D_b are the corresponding non-zero eigenvalues of S_b. S_w is transformed to

$$K_w = D_b^{-1/2} Y^T S_w Y D_b^{-1/2} . \tag{5-5}$$

K_w is diagonalized by eigenanalysis,

$$U^T K_w U = D_w . \tag{5-6}$$

The LDA transformation matrix for classification is defined as,

$$W = Y D_b^{-1/2} U D_w^{-1/2} . \tag{5-7}$$

In direct LDA, the null space of S_b is first removed. It is assumed that the null space of S_b contains no discriminative information at all. This assumption is not true. In direct LDA, projection vectors are restricted in the subspace spanned by class centers. But the optimal discriminant vectors do not necessarily lie in the subspace spanned by class centers. This point can be clearly illustrated in the Figure 5-1. For a binary classification problem, using direct LDA, the derived discriminant projection vector is constrained to the line passing through the two class centers. But according to the Fisher criteria (2-10), the optimal discriminant vector should be the in direction of line B. Furthermore, direct LDA also encounters the singularity problem of S_w. To keep the information in the null

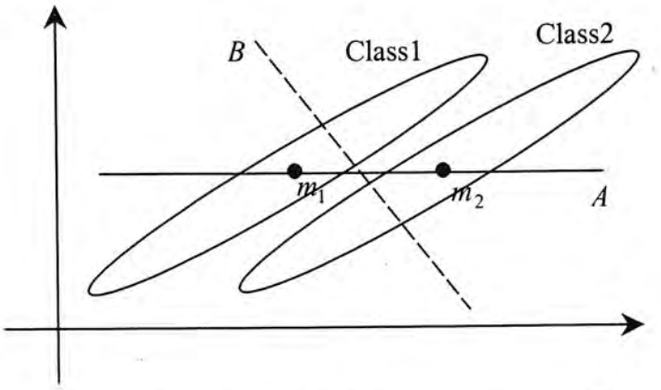


Figure 5-1 Direct LDA for a two-class problem.

Using direct LDA, the discriminant vector is constrained to the line A passing through the two class centers m_1 and m_2. But according to the Fisher criteria, the optimal discriminant projection should be line B

space of S_w, D_w has to contain zero eigenvalues. But in (5-7), the data is whitened by $D_w^{-1/2}$ for classification, and singularity will occur in this case.

5.1.3 Discussion

As discussed above, all these proposed LDA approaches have lost some discriminative information in the data space. This point can be further illustrated in Figure 5-2. **A** is the subspace spanned by the eigenvectors of S_w, and **B** is the subspace spanned by the eigenvectors of S_b. Since the total scatter matrix S_t is equal to the summarization of S_w and S_b [36],

$$S_t = S_w + S_b, \tag{5-8}$$

the face space is composed of **A** and **B**. When $\mathbf{B} \subseteq \mathbf{A}$ as shown in Figure 5-2 (a), LDA in the principal subspace of S_w can keep all the discriminative information in data space. When $\mathbf{A} \subseteq \mathbf{B}$ as shown in Figure 5-2 (b), direct LDA can keep all the discriminative information. When $\mathbf{A} \cap \mathbf{B} = \phi$ as shown in Figure 5-2 (c), LDA in the null space of S_w can keep all the discriminative information. But when **A** and **B** are only partially intersected as shown in Figure 5-2 (d), some discriminative information will definitely be lost using the conventional LDA approaches.

Furthermore, conventional LDA approaches suffer from the problem of overfitting. LDA vectors are tuned to the training set with the existence of noise. As suggested in [24], an eigenvector will be very sensitive to small perturbation if its eigenvalue is close to another eigenvalue of the same matrix. The eigenvectors of S_w with very small eigenvalues are unstable. They may contain discriminative

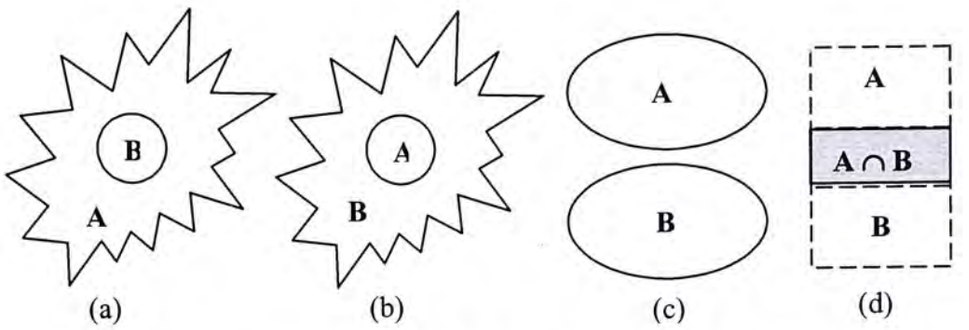


Figure 5-2 Analysis different LDA approaches.

A is the subspace spanned by the eigenvectors of S_w, **B** is the subspace spanned by S_b, and **A** ∪ **B** is the whole data space. In case (a), **B** ⊆ **A**, LDA in the principal space of S_w can keep all the discriminative information. In case (b), **A** ⊆ **B**, direct LDA can keep all the discriminative information. In case (c), **A** ∩ **B** = ϕ, LDA in the null space of S_w can keep all the discriminative information. In case (d), **A** and **B** are only partially intersect, and so discriminative information in data space will definitely be lost in conventional LDA approaches.

information, but also may be very sensitive to noise. In (2-13) and (5-8), LDA in the principal subspace of S_w and direct LDA all need to whiten with the inverse of eigenvalues of S_w. Some trivial eigenvalues are not well estimated because of the small sample size problem, but they can substantially change the LDA vectors. If data vector is whitened on noisy eigenvectors, overfitting will happen. For LDA in the null subspace of S_w, the rank of S_w, $r(S_w)$, is sensitive to noise. $r(S_w)$ is bounded by $\min(M-L, N)$, where M is the total training sample number, L is the class number, and N is the dimensionality of the data vector. $r(S_w)$ is almost equal to this bound because of the existence of noise. When the sample number is very large, the null space of S_w becomes very small, so much discriminative information outside this null space will be lost [38].

5.2 Discriminant analysis in dual intrapersonal subspaces

Bayesian face recognition could combine the discriminative features in transpersonal principal and complementary subspaces under a probabilistic model. Since Bayesian algorithm can be viewed as an intermediate step of LDA, this implies that LDA can be simultaneously applied in the principal and null subspaces of S_w to make full use of the discriminative information in data space, and the two parts of discriminative features can be combined under a probabilistic model. Based on this, we develop a novel discriminant analysis method. At the training stage,

1. Compute the within-class scatter matrix S_w and between-class scatter matrix S_b from the training set.
2. Apply eigenanalysis to S_w, and compute the principal subspace F, with K eigenvectors $V = [\phi_1, ..., \phi_K]$, and its complementary subspace $\overline{F}$. Estimate the average eigenvalue ρ in $\overline{F}$.
3. All of the class centers are projected to F and normalized by eigenvalues. The between-class scatter matrix S_b is transformed to

$$K_B^P = \Lambda^{-1/2} V^T S_b V \Lambda^{-1/2}, \tag{5-9}$$

where Λ is the eigenvalue matrix for F. Apply eigenanalysis to the transformed between-class scatter matrix K_b^P, and compute l_P eigenvectors Ψ_P with the largest eigevalues. The l_P discriminative vectors in F are defined as

$$W_P = V \Lambda^{-1/2} \Psi_P. \tag{5-10}$$

4. Project all the class centers to F and compute the reconstruction difference as

$$\begin{aligned} A_r &= A - VV^T A \\ &= \left(I - VV^T\right) A \end{aligned} \tag{5-11}$$

where $A = [\bar{m}_1, ..., \bar{m}_L]$ is the class center matrix. In fact, A_r is the projection of A into $\overline{F}$. In $\overline{F}$, the between-class scatter matrix is transformed to

$$K_B^C = \left(I - VV^T\right) S_B \left(I - VV^T\right) \tag{5-12}$$

Apply eigenanalysis to the transformed between-class scatter matrix K_b^C, and compute l_c eigenvectors Ψ_C of K_B^C with the largest eigenvalues. The l_C discriminative vectors in the $\overline{F}$ are defined as

$$W_C = \left(I - VV^T\right) \Psi_C \tag{5-13}$$

For recognition,

1. All the prototype faces $\{x_j\}$ in the gallery are projected to the discriminant vectors in F to get

$$\bar{a}_j^P = W_P^T x_j. \tag{5-14}$$

They are also projected to the discriminant vectors in $\overline{F}$ to get

$$\bar{a}_j^C = W_C^T x_j. \tag{5-15}$$

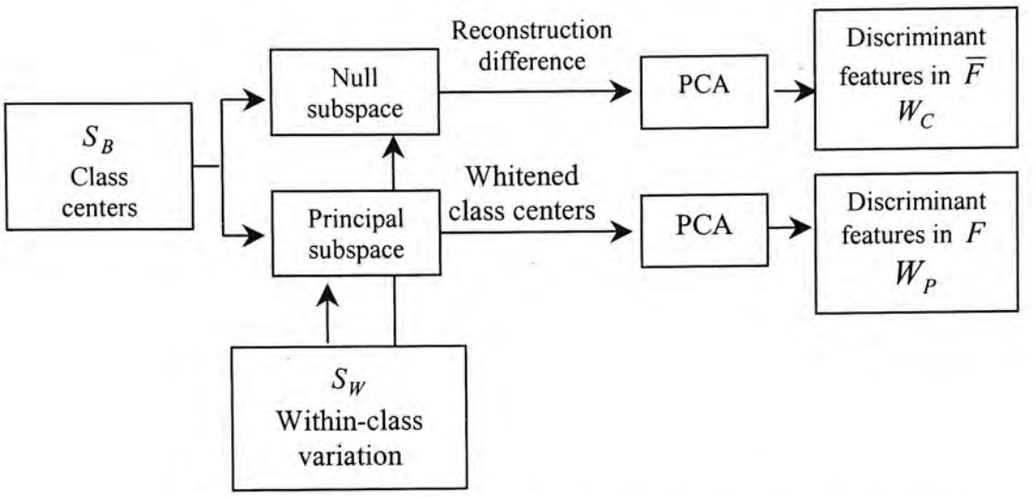


Figure 5-3 Training discriminant vectors in dual intrapersonal subspaces.

$\{\vec{a}_j^P, \vec{a}_j^C\}$ are stored as low dimensional features.

2. At the runtime, when a data vector x_t is input, it is also projected to the discriminant vectors in the dual subspaces to get low dimensional feature vectors

$$\vec{a}_t^P = W_P^T x_t \tag{5-16}$$

$$\vec{a}_t^C = W_C^T x_t \tag{5-17}$$

3. Class is found to minimize the distance measure as

$$d(\Delta) = \left\|\vec{a}_j^P - \vec{a}_t^P\right\|^2 + \left\|\vec{a}_j^C - \vec{a}_t^C\right\|^2 / \rho \tag{5-18}$$

The method of discriminant analysis in dual subspaces has several advantages to conventional LDA and Bayes methods.

(1) In this approach, LDA is generalized to take advantage of discriminative information in the full face space, while other LDA approaches all lose some distinctive information. As discussed in Section 5.1, in principal and null subspaces of S_w, LDA vectors are computed using different criterions,

$$\begin{cases} W_P = \arg\max \dfrac{\left|W_P^T S_b W_P\right|}{\left|W_P^T S_w W_P\right|} \\ W_P^T S_w W_P \neq 0 \end{cases} \tag{5-19}$$

$$\begin{cases} W_C = \arg\max \left|W_C^T S_b W_C\right| \\ W_C^T S_w W_C = 0 \end{cases} \tag{5-20}$$

Both of the two parts of discriminative features are effective for classification. In F, after whitening, the intrapersonal variation has been effectively reduced, such that $W_P^T S_w W_P = I$. In $\overline{F}$, most of the intrapersonal variation has been removed, so

we have $W_P^T S_w W_P \approx 0$. Both the discriminative features contribute to the class centers in the two subspaces. However they cannot by directly combined,

$$d = \left\| W_P^T \left(x - x_j \right) \right\|^2 + \left\| W_C^T \left(x - x_j \right) \right\|^2 . \tag{5-21}$$

The principal subspace of S_w has been whitened, so projection vectors in W_P are not orthonormal. The two subspaces have different metric scales. It is unreasonable, at least not optimal, to combine the distances in the two subspaces directly. Using Eq. (5-18), we develop a much better way to combine the two parts of features. In this approach, the null space of S_w is also whitened by the average residue eigenvalue. In (5-18), $\left\| \bar{a}_j^P - \bar{a}_t^P \right\|^2$ and $\left\| \bar{a}_j^C - \bar{a}_t^C \right\|^2 / \rho$ computed under the same metric scale measure the distances in two subspaces equally whitened by the eigenvalue spectrum of S_w.

(2) It is more stable and insensitive to noise than other LDA approaches. The eigenvectors of S_w with very small eigenvalues are unstable and sensitive to small perturbation. This approach avoids computing these unstable eigenvectors by grouping them into the complementary subspace to encode discriminative information. The eigenvalue spectrum of S_w is better estimated, and it avoids whitening with very small eigenvalues.

(3) This approach is also an improvement to the Bayesian face recognition. It is more effective for classification and efficient in computation compared with Eq (3-3). Besides effective reducing the intrapersonal variation like the Bayes, it further separates class centers, and removes some noise disturbance by compacting the discriminative features. Computing the reconstruction error $\varepsilon^2(x)$ in (3-3) is expensive. Its computation cost is comparable to the correlation between two high dimensional data vectors. Although an efficient computation is proposed in [10], by computing the Mahalanobis distance in F, it is at the cost of discarding the information in $\overline{F}$. Our approach is much more effective, since it only needs to compute the distances of $l_p + l_c$ features. LDA can be interpreted under a probabilistic model. When each class has the same sample number n, the intrapersonal likehood estimated from LDA features,

$$\hat{P}(x_t \mid X_j) = \left[\frac{\exp\left(-\frac{1}{2} \left\| \bar{a}_t^P - \bar{a}_j^P \right\|^2 \right)}{(2\pi)^{K/2} \prod_{i=1}^{K} \lambda_i^{1/2}} \right] \left[\frac{\exp\left(-\frac{1}{2\rho} \left\| \bar{a}_t^C - \bar{a}_j^C \right\|^2 \right)}{(2\pi\rho)^{(N-K)/2}} \right], \tag{5-22}$$

is close to the optimal approximation to (2-20) using low dimensional features. This can be shown by Lemma 1.

Lemma 1. When each class has the same sample number n, the distance $\left\|a_t^P - a_j^P\right\|^2$ in F is the optimal approximation to the Mahalanobis distance $\sum_{i=1}^{K} \frac{y_i^2}{\lambda_i}$ in F using l_P features; the distance $\left\|a_t^C - a_j^C\right\|^2$ in $\overline{F}$ is close to the optimal approximation to the reconstruction error $\varepsilon^2(x)$ using l_C features.

Proof. Let Δ_j^P be feature vectors of face difference Δ_j projected to F and whiten by eigenvalues. Then, $\sum_{i=1}^{K} \frac{y_i^2}{\lambda_i}$ is the norm of the feature vector Δ_j^P,

$$\sum_{i=1}^{K} \frac{y_i^2}{\lambda_i} = \left\|\Delta_j^P\right\|^2 \tag{5-23}$$

Apply PCA to the set $\left\{\Delta_j^P\right\}$. According to the Theorem 1, the covariance matrix is

$$\begin{aligned} C &= \sum_i \left(\Delta_j^P\right)\left(\Delta_j^P\right)^T \\ &= 2M * S_t^P \\ &= 2M * \left(S_w^P + S_b^P\right), \end{aligned}$$

where, S_w^P, S_b^P, and S_t^P are the within-class scatter matrix, between-class scatter matrix, and total scatter matrix when all samples are projected to F and whitened by eigenvalues. Since $S_w^P = I$,

$$C = 2M * \left(S_b^P + I\right). \tag{5-24}$$

So the l_p largest eigenvectors of C are equal to the l_p largest eigenvectors of S_b^P. $\left\|a_t^P - a_j^P\right\|^2$ is computed from the l_p most dominant axes for the distribution of $\left\{\Delta_j^P\right\}$, so it is the optimal estimation to $\sum_{i=1}^{K} \frac{y_i^2}{\lambda_i}$ using l_p features.

Let Δ_j^C be feature vectors of face difference Δ_j projected onto $\overline{F}$,

$$\varepsilon^2(x) = \left\|\Delta_j^C\right\|. \tag{5-25}$$

In the similar way, the covariance matrix of set $\left\{\Delta_j^C\right\}$ is,

$$C = 2M * \left(S_b^C + S_w^C\right), \tag{5-26}$$

where S_w^C and S_b^C are the within-class scatter matrix and between-class scatter matrix in $\overline{F}$. S_w^C is close to zero in $\overline{F}$ (when $\overline{F}$ is exactly the null space of S_w, S_w^C is definitely zero), so S_b^C is almost equal to C. The l_C largest eigenvectors of S_b^C are almost equal to the l_C largest eigenvectors of C. So $\left\|a_i^C - a_j^C\right\|^2$ is close to the optimal approximant to $\varepsilon^2(x)$ using l_C features.

Since the Mahalanobis distance characterizes the intrapersonal likehood, from Lemma 1, we can get the following conclusion. When each class has the same sample number, the intrapersonal likehood estimated from LDA features are close to the optimal approximation to (2-20) estimated by probabilistic visual model using low dimensional features.

5.3 Experiment

In this section, we apply this novel approach to face recognition and compare it with conventional LDA approaches and the Bayesian face recognition by experiments on the data sets from the FERET face database and the XM2VTS database. Three conventional LDA approaches are selected for comparison: Fisherface, LDA in the null space of S_w, and direct LDA. Preprocessing for face image is similar to Section 4.

5.3.1 Experiment on FERET face database

The data set for experiment is similar to that of Section 4.1. Among the selected 1195 persons from the FERET database, 495 persons are used for training, and the remaining 700 persons are used for testing. For each testing person, one face image is in the gallery and the other is for probe.

First, we compare the new method with the Bayesian face recognition. Figure 5-4 reports their recognition accuracies with different feature number. The feature number for the new method is the summation of discriminative feature numbers in F and $\overline{F}$. In this experiment, we set l_P equal to l_C. The feature number for the Bayesian algorithm is the dimensionality (K) of F. Three similarity measures, DIFS, DFFS, and ML (DIFS+DFFS), for the Bayes are evaluated. But the feature number only affects DIFS on computation cost. Even for small K, DFFS and ML require high computational cost, since they need to compute the reconstruction

error. When the feature number is small, the recognition accuracy of DIFS is low, because too much discriminative information is lost in the complementary subspace. The new method outperforms DIFS at the same computational cost. It can achieve above 96% recognition accuracy, while the best performance for the three Bayesian distance measures is about 93%. This clearly demonstrates the superiority of the new method over the Bayesian algorithm. The improvement is due to that the new method utilizes the discriminative information in two subspaces and further extracts the face intrinsic difference based on class centers.

The new method also outperforms conventional LDA approaches. Figure 5-5 reports the accumulative scores comparison with Fisherface, LDA in the null space of S_w and direct LDA. The novel method has reduced 50% error rate than conventional approaches. The performance of LDA approaches is affected by the size of spaces spanned by S_w and S_b. Figure 5-6 and Table 5-1 report the recognition accuracies of the four LDA approaches, selecting different number of people as training set. When only a small number of people are selected for training, e.g. 50, the spaces spanned by S_w and S_b are small. Fisherface and direct LDA have very low recognition accuracies of 70%, because too much discriminative information has been discarded when removing the null spaces of S_w and S_b in the first step. LDA in the null space of S_w is better in this case. As the training set increases, the performance of Fisherface and direct LDA improves, since space dimensionalities of S_w and S_b increase, and much more discriminative information is included. But the accuracy for the null space of S_w significantly drops when 495 people are selected for training, because the null space is much smaller for large training set. The new approach is barely affected by the size of training set, and almost achieves the same high accuracies in different cases. This further proves that the new approach integrates the advantages of other LDA approaches.

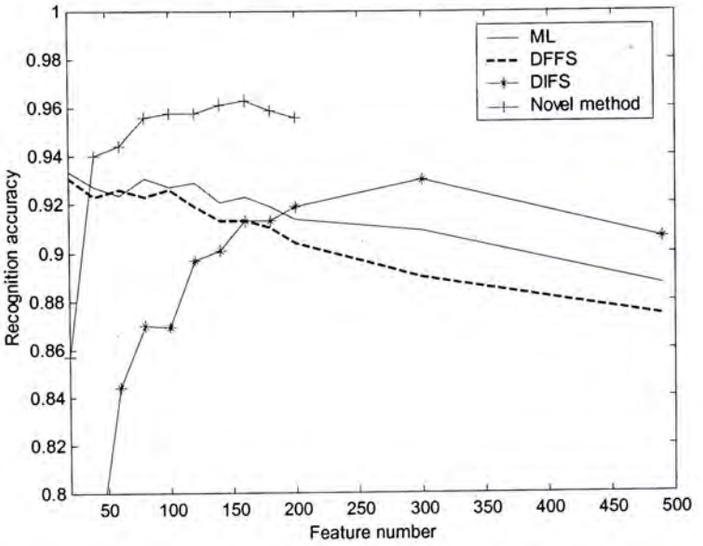


Figure 5-4 Recognition accuracy comparison of the new method with Bayesian face recognition.

Feature number for new method is the summation of feature numbers in principal subspace and complementary subspace. The feature number for Bayes (DIFS, DFFS, and ML) is the dimensionality of the intrapersonal principal subspace.

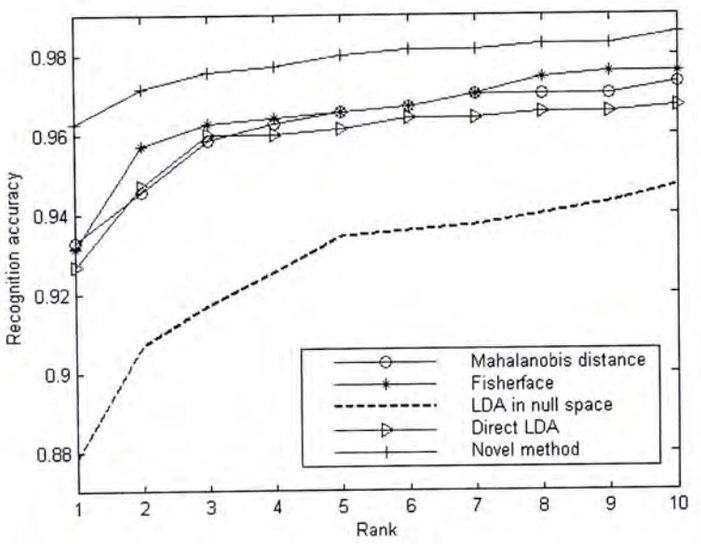


Figure 5-5 Accumulative scores for the new method, Bayes (Mahalanobis distance), Fisherface, LDA in null space, and direct LDA.

Table 5-1 Recognition accuracies of Fisherface, LDA in null space, direct LDA, and the new method using different numbers of persons for training on the FERET database.

	50	100	300	495
Fisherface	0.7743	0.8827	0.9271	0.9314
LDA in Null space	0.8800	0.9200	0.9271	0.8786
Direct LDA	0.7229	0.8357	0.9141	0.9270
New Method	0.8740	0.9143	0.9500	0.9629

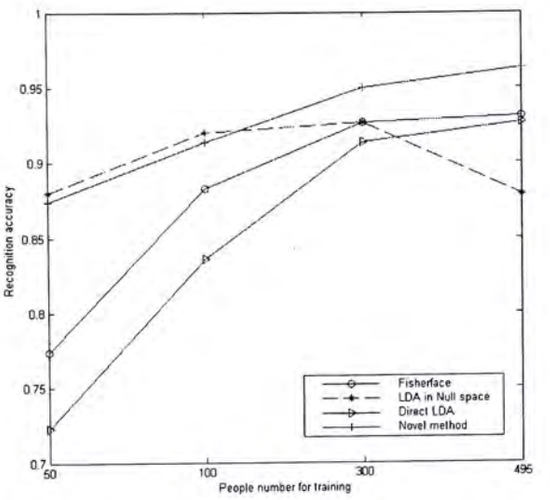


Figure 5-6 Recognition accuracies of Fisherface, LDA in null space, direct LDA, and the new method using different number of persons for training on the FERET database.

5.3.2 Experiment on the XM2VTS database

The data set from the XM2VTS database contains 295 people with 4 face images for each person. We use a cross-validation analysis for testing. The 1180 face images are partitioned into 4 folders. Each folder contains one face image for each individual. For each experimental trial, one folder is chosen as the probe set, and the remaining three folders are used as the gallery and training set. The recognition accuracies of the new method and other conventional methods on the four experimental trials and their mean accuracies are reported in Table 5-2. The results again clearly demonstrates the effectiveness of the new method.

Table 5-2 Face recognition accuracies on the four experimental trials of the data set from the XM2VTS database.

	Bayes (ML)	FisherFace	LDA in null space	Direct LDA	New method
1	0.9525	0.9593	0.9695	0.9085	0.9898
2	0.9627	0.9492	0.9627	0.9017	0.9864
3	0.9492	0.9661	0.9797	0.9356	0.9898
4	0.9695	0.9661	0.9763	0.9288	0.9898
Mean	0.9585	0.9602	0.9721	0.9186	0.9890

Chapter 6

Eigentransformation: Subspace Transform

In our framework, it is assumed that the intrapersonal variation is modeled as Gaussian distribution, $\tilde{T}$ and $\tilde{I}$ can be separated by PCA on the intrapersonal difference set. However, when sometimes $\tilde{T}$ is significant, this assumption may break down. The face difference even cannot be modeled as linear composition of $\tilde{I}$, $\tilde{T}$, and $\tilde{N}$. An extreme example is using face sketch to recognize face photo. Since face photo and sketch are in different modality, it is very difficult to directly recognize sketch from photos using conventional subspace methods. We can think the two kinds face images are in different subspaces. To alleviate this difficulty, we develop an eigentransformation algorithm [86] to transform different "stylistic" face images into one modality, thus the significant transformation difference $\tilde{T}$ can be effectively reduced. It can be viewed as the transformation between different face subspaces. In this thesis, eigentransformation is successfully applied to face sketch recognition. It is also found that it can be used for hallucination, rendering high-resolution face image from the low-resolution one. The two particular applications are described in this chapter.

6.1 Face sketch recognition

An important application of face recognition is to assist law enforcement. Automatic retrieval of photos of suspects from police mug-shot database can help the police narrow down potential suspects quickly. However, in most cases, the photo image of a suspect is not available. The best substitute is often a sketch drawing based on the recollection of an eyewitness. Therefore, automatically searching through a photo database using a sketch drawing is very useful. It will not only help the police to locate a group of potential suspects, but may also help the witness and the artist to modify the sketch drawing of the suspect interactively based on the similar photos retrieved.

However, due to the great difference between sketches and photos, and the unknown psychological mechanism of sketch generation, face sketch recognition

(a) Photo

(b) Sketch

Figure 6-1 Examples of photo-sketch pairs.

is much more difficult than the normal face recognition based on photo image. During the past three decades, many face recognition techniques have been proposed, however, few effective face sketch recognition systems can be found in previous researches. Methods using traditional photo-based face recognition techniques such as the eigenface method [63] and the elastic graph matching method [81] have been tested with very small sketch datasets. In [63], the sketch was normalized in geometry and blurred by a Gaussian filter in preprocessing, and then recognized by the eigenface method. The experiment includes only 7 sketches. The method in [81] recognized face sketches using Elastic Graph Matching [40], and was tested on 13 sketches only.

Photo and sketch have different modalities as shown by some samples in Figure 6-1. The key for sketch-based face photo recognition is to reduce the difference between the two modalities. In this thesis, we develop a sketch synthesis method based on separate transformation of photo texture and shape. This method significantly reduces the difference between photo and sketch. We show that the synthesized sketch by the separate transformation is a good approximation to the real one when the transformation procedure can be approximated as linear. A Bayesian classifier combining texture and shape features is then designed to recognize the probing real sketch from the synthesized pseudo-sketches. To

evaluate face sketch recognition performance on a large database, we construct a database containing photo-sketch pairs of 606 people. Experiments show that our method is much more effective than using the convectional photo-based methods directly. The new method is also shown to outperform human beings.

6.1.1 Eigentransformation

It is difficult to directly match photo and sketch since they are in different modalities. The starting point of our algorithm is to transform photo into sketch, so that recognition can be performed in the same modality. The relationship between photo and sketch is learnt from a set of training photo-sketch pairs using an eigentransformation procedure.

6.1.1.1 Algorithm

Face image can be reconstructed from eigenfaces in the PCA representation. Since eigenface is computed from the training set, we can show that the reconstructed face image can also be expressed as the linear combination of training samples.

We represent the photo training set by an N by M matrix, $[\bar{P}_1, \bar{P}_2, \ldots, \bar{P}_M]$, where $\bar{P}_i$ is the photo vector, N is the number of image pixel, and M is the number of training samples. In PCA, a set of eigenvectors $E_p = [e_1, \ldots, e_K]$, also called eigenfaces, are computed from the covariance matrix,

$$C = \sum_{i=1}^{M} (\bar{P}_i - \bar{m}_p)(\bar{P}_i - \bar{m}_p)^T = A_p A_p^T, \tag{6-1}$$

where $\bar{m}_p$ is the photo mean face, and A_p is the photo sample matrix,

$$A_p = [\bar{P}_1 - \bar{m}_p, \ldots, \bar{P}_M - \bar{m}_p] = [\bar{P}'_1, \ldots \bar{P}'_M]. \tag{6-2}$$

According to the singular value decomposition theorem, E_p also can be computed from,

$$E_p = A_p V_p \Lambda_p^{-1/2}, \tag{6-3}$$

where V_p and Λ_p are the eigenvector and eigenvalue matrix for $A_p^T A_p$.

For a new face photo $\bar{P}$, it can be reconstructed from the eigenfaces by,

$$\bar{P}_r = E_p \bar{w}_p + \bar{m}_p, \tag{6-4}$$

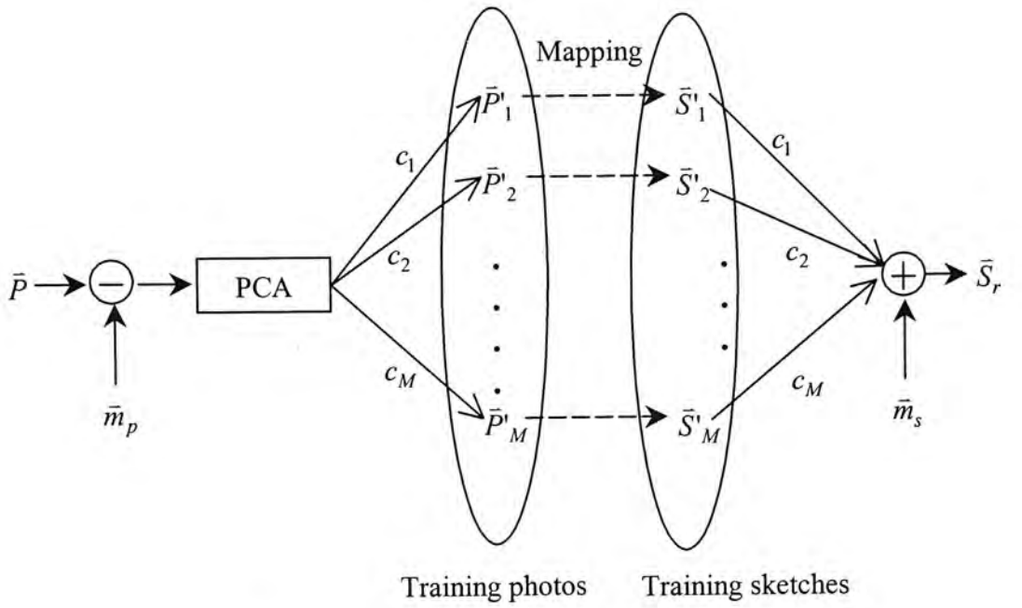


Figure 6-2 Eigentransformation procedure.

where $\vec{w}_p$ is the weight vector computed by projecting the face photo onto the eigenfaces,

$$\vec{w}_p = E_p^T\left(\vec{P} - \vec{m}_p\right). \tag{6-5}$$

From (6-3) and (6-5), the reconstructed photo can be represented by

$$\vec{P}_r = A_p V_p \Lambda_p^{-\frac{1}{2}} \vec{w}_p + \vec{m}_p = A_p \vec{c} + \vec{m}_p, \tag{6-6}$$

where $\vec{c} = V_p \Lambda_p^{-1/2} \vec{w}_p = [c_1, \ldots, c_M]^T$. Equation (6-6) can be rewritten as,

$$\vec{P}_r = \sum_{i=1}^{M} c_i \vec{P}'_i + \vec{m}_p. \tag{6-7}$$

This shows that the reconstructed photo is an optimal approximation to the original face photo using a linear combination of the M training photos. Replacing each training photo with its corresponding sketch $\vec{S}'_i$, and replacing the mean photo $\vec{m}_p$ with the mean sketch $\vec{m}_s$, we get,

$$\vec{S}_r = \sum_{i=1}^{M} c_i \vec{S}'_i + \vec{m}_s. \tag{6-8}$$

$\vec{S}_r$ is the synthesized sketch expected to resemble the real sketch. The eigentransformation procedure is shown in Figure 6-2.

6.1.1.2 Linear assumption

Two conditions are needed for the eigentransformation to work:

- A new face can be reconstructed from training samples by PCA.

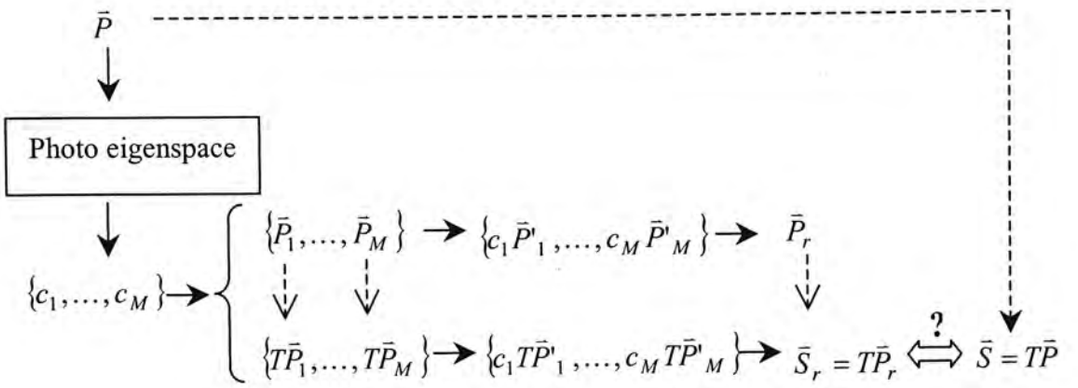


Figure 6-3 Eigentransformation with the assumption that the transformation between photo and sketch is linear.

We use the dotted arrow to represent the drawing process

- The transformation between photo and sketch can be approximated as a linear process.

Assuming that there is a linear transformation matrix T that can ideally transform a photo to sketch, the transformation can be expressed as

$$\vec{S} = T\vec{P}\,. \tag{6-9}$$

For the training set, we have

$$\vec{S}'_i = T\vec{P}'_i\,, \tag{6-10}$$

$$\vec{m}_s = T\vec{m}_p\,. \tag{6-11}$$

As shown in Figure 6-3, when $\vec{P}$ is projected onto the photo eigenspace, a group of coefficients $\{c_1,c_2,\ldots,c_M\}$ on the training set can be obtained, from which $\vec{P}_r$ and $\vec{S}_r$ can be reconstructed by the linear combination of training photos and training sketches respectively. From Eq. (6-7) and (6-8), replacing $\vec{S}'_i$ and $\vec{m}_s$ with (6-10) and (6-11), we have

$$\vec{S}_r = \sum_{i=1}^{M} c_i T\vec{P}'_i + T\vec{m}_p = T\left(\sum_{i=1}^{M} c_i\vec{P}'_i + \vec{m}_p\right) = T\vec{P}_r\,. \tag{6-12}$$

This shows that the reconstructed sketch is in fact a sketch drawn based on $\vec{P}_r$. Photo-based face recognition studies [59] have shown that for eigenface reconstruction $\vec{P}_r$ is close to $\vec{P}$ because of the facial structural similarity. Therefore, comparing Eq. (6-9) and (6-12), we see that the reconstructed sketch $\vec{S}_r$ should be similar to $\vec{S}$. The linear requirement is critical for Eq. (6-12) to hold. The linear assumption is not unreasonable since some highpass-filered images are actually sketch-like. For a simple example, the edge gradient map obtained by linear edge detector can be seen as a line drawing sketch. Thus it is possible to use

linear operator to generate sketch-like images from the original photo. Of course, for the real sketch drawn manually, the transform cannot be strictly linear, but an approximation.

Even though the transformation may be simplified to be linear, the transformation matrix T is still too complicated to be expressed explicitly. For an image of size 128 by 128, the length of the image vector is 128^2. So the matrix T has 128^4 elements to be defined. Eigentransformation takes advantage of the linear property and face structural similarity to generate the sketch by using only a small number of training samples without actually deriving the large transformation matrix.

6.1.2 Sketch synthesis

Since the performance of the sketch synthesis by eigentransformation depends on the linear assumption. Given the fairly complex structure of human face, this assumption is rather difficult to be demonstrated. Since the difference between sketch and photo exists in both texture and shape, corresponding points in the sketch and photo can be quite different without proper alignment, thus the process becomes difficult to be described by a linear process. However, if we separate the texture and shape, and then treat them independently, a closer linear correspondence can then be established.

The shape distortion is somewhat caused by that the artist tries to exaggerate some distinctive features just like caricature. For example, if a face has a big nose in a photo, the nose drawn in the sketch will be even bigger. A study in [26] suggested that the shape exaggeration could be approximated as,

$$G_s = E\left(G_p - G_p^m\right) + G_p, \tag{6-13}$$

where G_p is the photo shape vector, G_s is the sketch shape vector, G_p^m is the mean photo shape, and E is the exaggeration matrix. The difference between the photo shape and the mean shape is exaggerated. Averaging both sides of Eq. (6-13), we get the mean shape of sketch,

$$G_s^m = E\left(G_p^m - G_p^m\right) + G_p^m = G_p^m. \tag{6-14}$$

The mean shapes of photo and sketch are actually the same. Subtract Eq. (6-13) from Eq. (6-14), we have,

$$\left(G_s - G_s^m\right) = (E + I)\left(G_p - G_p^m\right) \tag{6-15}$$

Thus we prove that the shape transformation between photo and sketch can be approximated as linear.

The texture in a sketch is formed by the grayscale changes in small local areas. It is reasonable to assume that the grayscale around a fiducial point in the sketch is mainly influenced by the grayscale around the same fiducial point in the photo. However, because of shape distortion, the same fiducial points have different coordinates in different photos and sketches, thus it is difficult to derive the linear relation for the local texture transformation. Therefore it is necessary to separate the shape from texture.

We represent face shape with a graph containing the coordinates of a set of fiducial points. A mean shape is computed from the training set. In order to remove shape factor, we warp the face image to the mean shape using the affine interpolation based on a set of triangles. After alignment, the fiducial points in different face photos and sketches finally correspond to the same position. We observe that the sketch grayscales after shape alignment also has a similar style of exaggeration as the sketch shape as shown by Eq. (6-13). If an area in photo is light color, the artist will leave it blank in the sketch; if an area is relatively dark, the artist tends to emphasize it more with shade texture. Therefore, at least within a small local neighbor, there is a linear trend, thus a linear relation similar to Eq. (6-15) can be derived for texture transform. Of course, this is a very rough approximation, since an artist will not decide on the grayscale of a small area only based on the grayscales of the same area in the photo. For precise description of the texture transformation, the whole picture has to be taken into consideration.

Finally, the sketch synthesis system based on separate shape and texture eigentransformation can be implemented through the following steps, as shown in Figure 6-4:

- For an input face photo P, locate all the fiducial points on the face graph model to extract shape information.
- Warp the face image to a mean face shape derived from training set to separate the texture I_p and shape G_p from the photo image.
- Apply eigentransformation to the photo texture and shape respectively to generate texture I_s and shape G_s for the sketch.

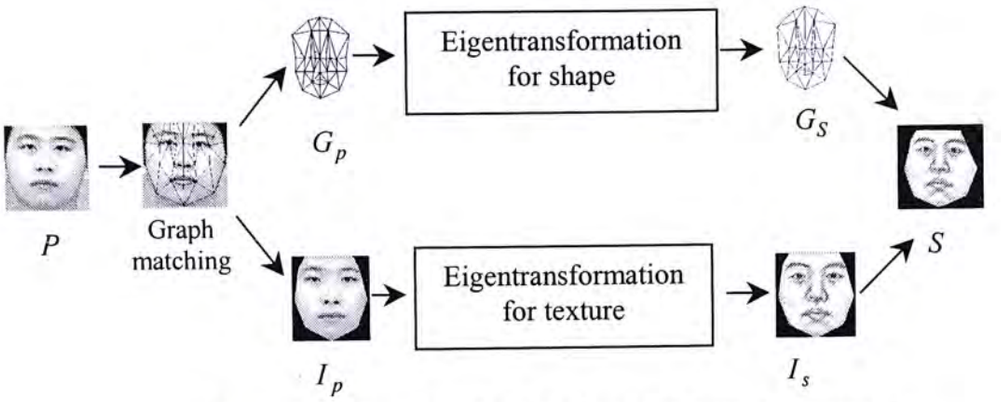


Figure 6-4 Framework of the face sketch synthesis system.

- Warp the generated texture from the mean shape to the sketch shape to produce the final synthesized sketch S.

6.1.3 Face sketch recognition

Face sketch recognition is based on the matching between the probing real sketch and the synthesized pseudo-sketch from photo. In this section, we present the PCA and Bayesian classifiers for recognition. For classification, we extract a set of salient geometric measures from the face graph to represent the shape feature, including the sizes and relative positions of nose, eyes, eyebrows, and face contours etc., and the texture vector is normalized by the shape.

6.1.3.1 PCA classifier

Let $G \in R^{N_1}$ and $I \in R^{N_2}$ represent the shape and texture vectors, where N_1 and N_2 are the vector length for shape and texture. The feature vectors used here is similar to the features used for photo-based recognition in active shape models [74]. Eigenspaces for shape and texture are computed from the sketch training set. In the PCA classifier, feature vectors are projected to eigenspaces to get the low dimensional features,

$$\bar{x} = E_G(G - m_G), \tag{6-16}$$

$$\bar{y} = E_I(I - m_I), \tag{6-17}$$

where E_G and E_I are the eigenvector matrices of shape and texture respectively, and m_G and m_I are the averages of shape and texture respectively. The shape and texture features are normalized to unit norms, and form an integrated feature,

$$\bar{z} = \left(\frac{\bar{x}^T}{\|\bar{x}\|} \frac{\bar{y}^T}{\|\bar{y}\|} \right)^T. \tag{6-18}$$

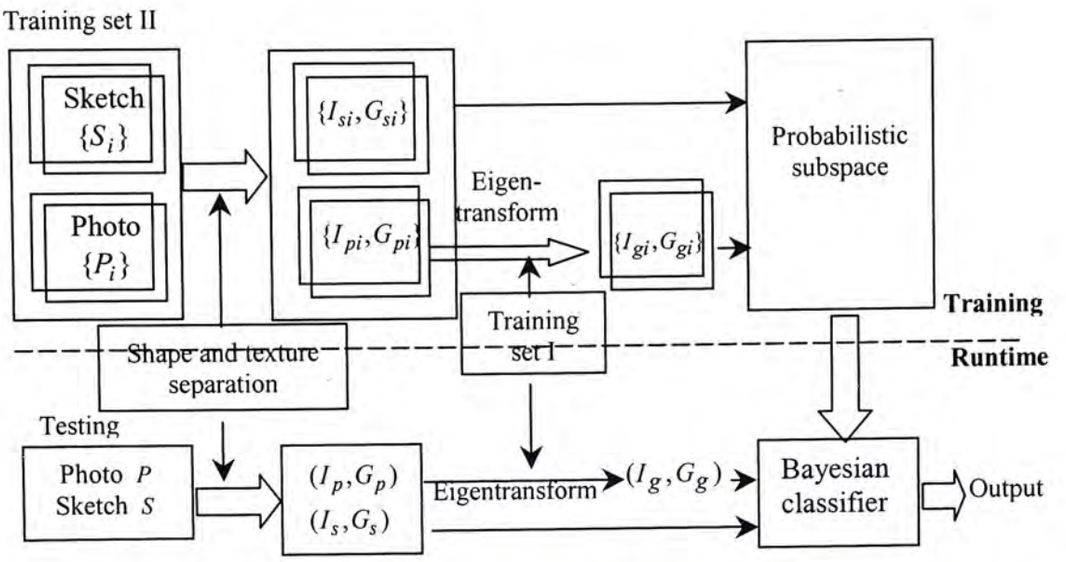


Figure 6-5 Face sketch recognition using eigentransformation and the Bayesian classifier.

Classification is based on the Euclid distance,

$$d = \| \bar{z}_s - \bar{z}_g \| . \tag{6-19}$$

where $\bar{z}_s$ and $\bar{z}_s$ are the integrated features for probe sketch and pseudo-sketch from photo.

6.1.3.2 Bayesian Classifier

Although eigentransformation let the matching be performed in the same modality, the synthesized sketch is still not a perfect estimation to the real one. To further reduce the effect of transformation error at recognition stage we use the Bayesian classifier in this section. Δ is defined as the difference between the real sketch and synthesized sketch.

In our algorithm, we separate the face image into shape and texture, and assume that they are independent. The Bayesian classifier is modified to integrate the two kinds of information by,

$$\hat{P}(\Delta \mid \Omega_I) = \hat{P}((\Delta_I, \Delta_G) \mid \Omega_I) = \hat{P}(\Delta_I \mid \Omega_I) \cdot \hat{P}(\Delta_G \mid \Omega_I). \tag{6-20}$$

Δ_I and Δ_G are the face difference in texture and shape. where $\hat{P}(\Delta_I \mid \Omega_I)$ and $\hat{P}(\Delta_G \mid \Omega_I)$ are the intrapersonal likehoods for texture and shape respectively.

Figure 6-5 describes the diagram of sketch recognition using eigentransformation and Bayesian classifier. The procedure is divided into two stages: training and runtime. There are two training sets. Training set I is for eigentransformation, and training set II is used to compute the probabilistic subspace for Bayesian classifier. At the training stage,

- Use training set I to compute the photo-to-sketch eigentransform coefficients.
- Photos and sketches in training set II are separated into shape and texture.
- Photo texture and shape $\{I_{pi}, G_{pi}\}$ in training set II are transformed to pseudo-sketches $\{I_{gi}, G_{gi}\}$.
- The texture and shape probabilistic subspaces are derived from training sketches $\{I_{si}, G_{si}\}$, and pseudo-sketches $\{I_{gi}, G_{gi}\}$.

At the runtime stage,

- The photo and sketch for matching are separated into texture and shape.
- The photo texture and shape (I_p, G_p) are transfromed to pseudo-sketch (I_g, G_g).
- Texture and shape features for sketch and pseudo-sketch are input to Bayesian classifier, and the face sketch is recognized.

6.1.4 Experiment

6.1.4.1 Sketch synthesis performance

The data set for sketch synthesis contains 188 persons. For each person, there is a face photo and two sketches drawn by different artists. We adopt the "leave-one-out" methodology. For each time, one person is selected for testing and the photos and sketches for the remaining 187 people are used as training set. Figure 6-6 gives some results of our sketch synthesis system. For each input face photo (a), (b) (d) give two generated sketches based on two kinds of sketch training sets drawn by different artists. The individual for testing is not in the training set. Our result is very similar to the real sketch on both texture and shape. In the generated sketch face, the skin color has been transformed to thesis texture, and there is noticeable shadow just like that added by pencil. The distinctive features on the face photo have been captured and exaggerated. The two artists have different drawing styles. The sketch drawn by artist A has a heavier shadow effect. The sketch drawn by artist B has lighter shadow and thinner lines, perhaps caused by a sharper pencil, and has a bigger exaggeration in shape. These stylistic differences caused by artists and drawing tools, can be noticeably exhibited on our generated sketches. Using different training sets, output of our system will involve different styles.

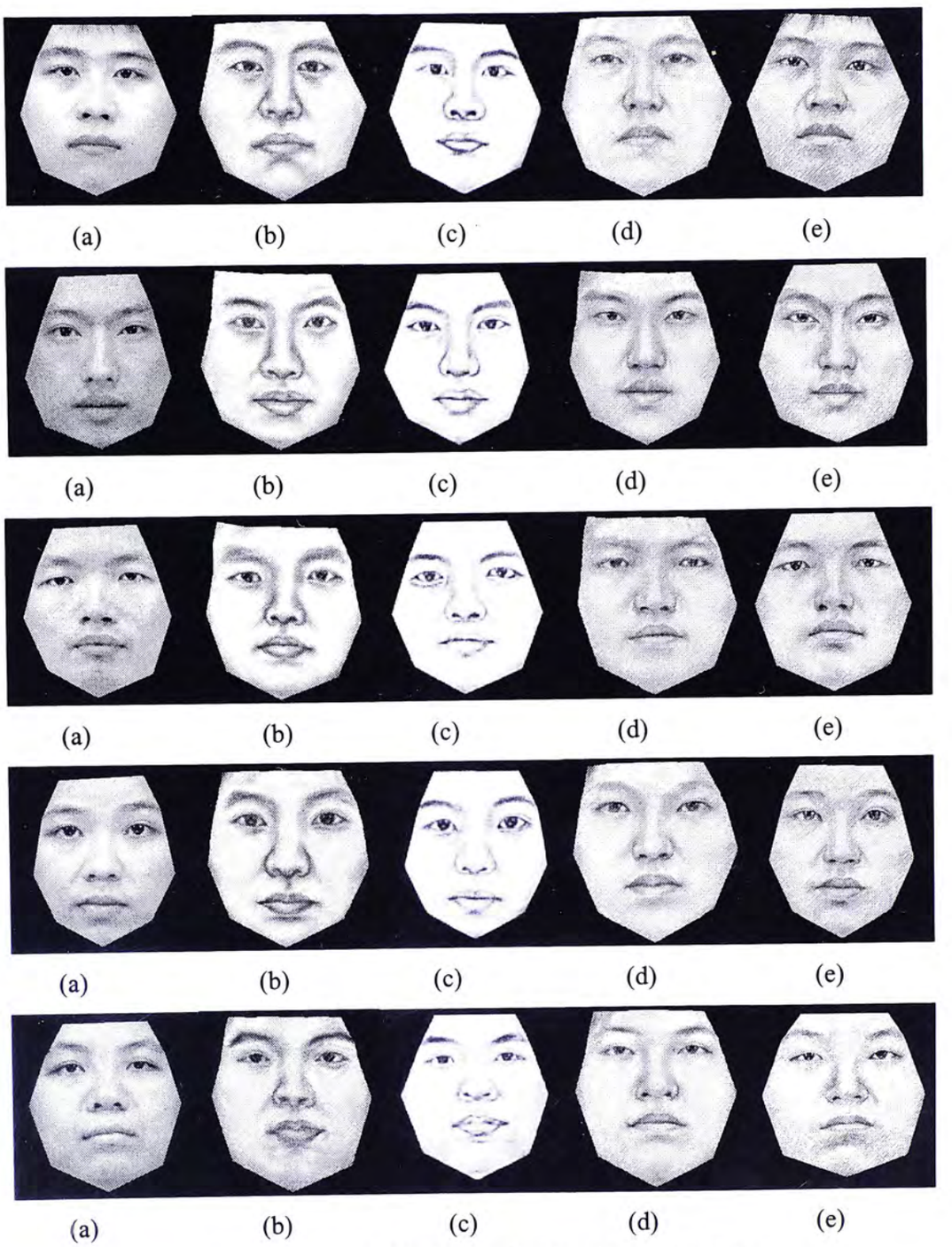
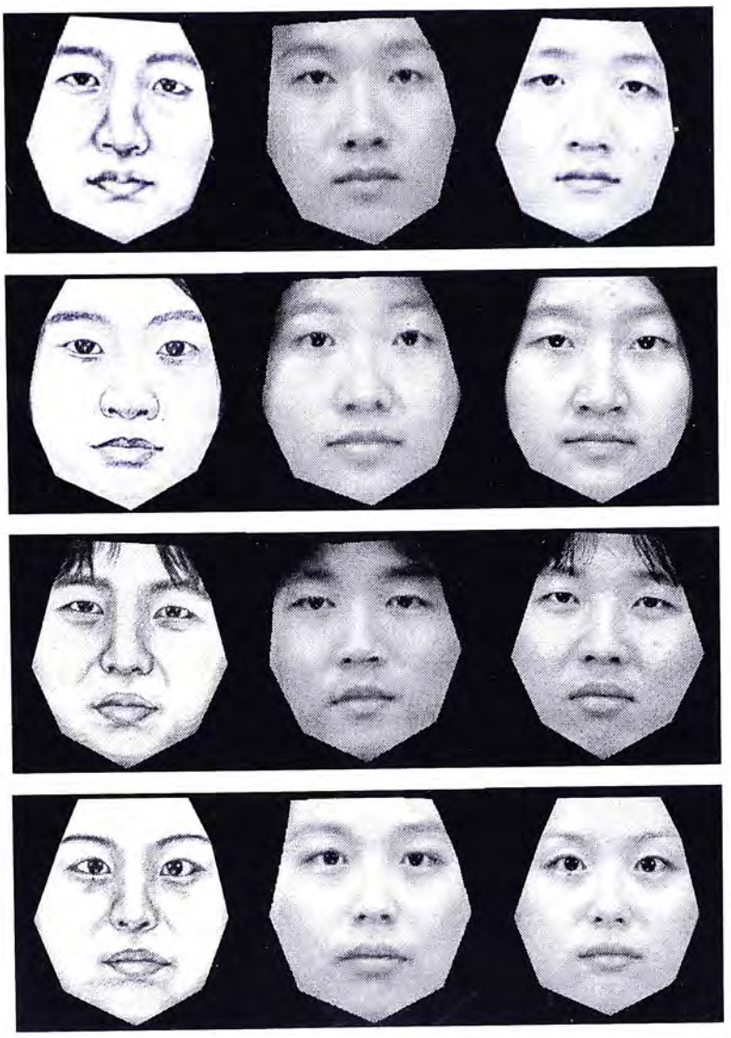


Figure 6-6 Facial sketch synthesis based on full face.

(a) is the input face photo, (b) is the generated sketch based on the training set drawn by artist A, (c) is the sketch drawn by artist A according to the input photo, (d) is the generated sketch based on the training set drawn by artist B, and (e) is the sketch drawn by artist B according to the input face photo.



(a) (b) (c)

Figure 6-7 Generate photo from the input sketch.

(a) is the input sketch drawn by the artist, (b) is the generated photo, and (c) is the real photo.

In an inverse procedure, our system can also generate a photo from the input sketch, just exchanging the positions of sketch and photo. Some results are shown in Figure 6-7. Contrary to sketch generation, which exaggerates features, the generated photo is similar with the real photo, and some distinctive features are de-emphasized, tending to the mean face.

Figure 6-8 shows the improvement of separating the face image into texture and shape. The first row is the results of applying eigentransformation directly on face images, which are just aligned by eye centers. The output images are blurred and have aliasing noise because of the non-linear difficulty. When texture and shape are separated, the results as shown in the second row, have clear and sharp appearance.

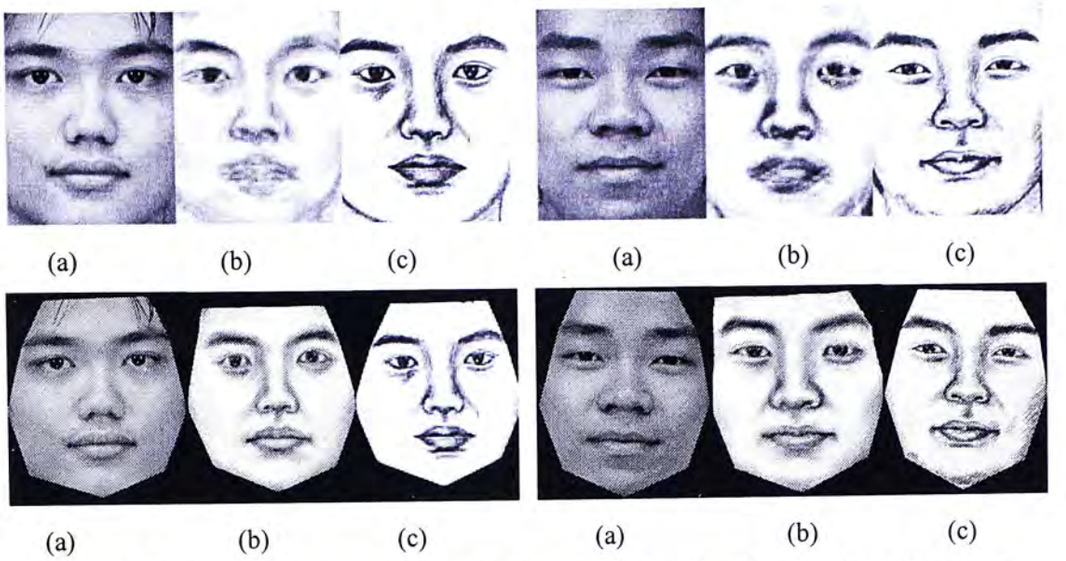


Figure 6-8 Comparison of the direct eigentransformation (first row) with separate transformation on texture and shape (second row).

(a) the input photo; (b) the synthesized sketch; (c) the sketch drawn by the artist.

6.1.4.2 Sketch recognition performance

To evaluate the face sketch recognition performance on large database, we construct a sketch database containing 606 people. For each people, there is a frontal photo face image, and a face sketch drawn by an artist. In this experiment, the 606 people are partitioned into three sets. Training set I and II contain 153 photo-sketch pairs each, and the testing set contains 300 photo-sketch pairs. Human hair is discriminative feature for short-term recognition, but it may vary significantly over a long period. We remove most of the hair and background in preprocessing.

Table 6-1 reports the sketch recognition accuracies using three different classifiers applied on four kinds of features, "copped face", texture, shape and the integration of texture and shape. The direct PCA method treats the probing sketch as a regular photo, and match photo and sketch in the eigenspace computed from the photo training set. The recognition performance is poor. The low accuracy on shape demonstrates that the reason for photo and sketch look alike is not because of the geometrical similarity of facial components.

The second classifier is PCA based on eigentransformation. The matching is performed between the probing sketch and the synthesized sketch in the eigenspace computed from the training sketches. It achieves significant improvement to the direct PCA method, since the match is performed in the same modality after transformation. The experiment also shows the improvement of

transforming face texture and shape separately and integrating them in recognition. It is much better than applying transformation on the "cropped face" without separating texture and shape. In the third classifier, the Bayesian algorithm further reduces the transformation error. The Bayesian classifier based on eigentransformation integrating texture and shape feature has the highest recognition accuracy 81.3%.

In Table 6-2, we compare our new method with two conventional face recognition method, eigenface [51] and Elastic Graph Matching (EMG) [40], both of which have been successfully applied to face photo recognition. As discussed in the introduction, both methods [63][81] have also been tested on very small datasets of sketches in previous study. The results in Table 6-2 clearly demonstrate the superiority of our algorithm over these conventional methods. Using a testing set containing 300 photo-sketch pairs, the first match for conventional methods is no more that 30%, and the tenth rank is no more than 60%. Our algorithm significantly improves the first match to 80%, and the tenth rank to 97%.

Table 6-1 Recognition accuracies using different features and classifiers (%).

	Cropped face	Texture	Shape	Texture + Shape
PCA	6.3	5.3	30.7	25.0
Eigentransform + PCA	53.7	45.0	35.3	75.0
Eigentransform + Bayes	74.3	56.7	53.0	81.3

Table 6-2 Acumulative match score for eigenface, EGM, and the novel method (%).

	1	2	3	4	5	6	7	8	9	10
Eigenface	6.3	8.0	9.0	9.3	11.3	13.3	14.0	14.0	14.3	16.0
EGM	25.3	32.3	40.0	43.0	46.7	48.7	53.0	54.3	56.3	57.7
New method	81.3	91.0	94.7	95.7	96.7	97.0	97.0	97.0	97.0	97.0

6.1.4.3 Comparison with human recognition

We conduct two experiments to compare the new method with sketch recognition by human beings. If we can demonstrate that automatic recognition by computers can perform better than human beings, we can then use computers to systematically conduct large-scale search in a large photo-ID database.

We select 100 photo-sketch pairs from the testing set for human recognition. Similar to automatic recognition, the hair is removed in the cropped faces. Thirty candidates are asked to do the test. In the first experiment, a sketch is shown to a human test candidate for a period of time, then the sketch is taken away before the photo search starts. The candidate tries to memorize the sketch, then go on to search the photo database without the sketch reference in front. The candidate can go through the database and are allowed to select up to 10 photos that are similar to the sketch. He can then rank the selected photos according to the similarity level to the sketch. This is closer to real application scenario. Since, people usually see the sketch of a criminal suspect in a newspaper or on TV briefly, then they have to rely on their memory to match the sketch with the suspect in real life.

For the second experiment, we allow the test candidate to look at the sketch while they search through the photo database. This simulates the case when an eye witness looks though the police database for a suspect.

The encouraging experimental results in Figure. 6-9 shows that a computer can perform better than a human being. The human performance for the first experiment is much lower. This is not only because of the difference between photo and sketch, but also because of the memory distortion, since it is difficult to precisely memorize the sketch. In fact, people are very good at distinguishing familiar faces, such as relatives and famous public figures, but are not very good at distinguishing strangers. Given the good automatic recognition results, we can now perform automatic searching of a large database using a sketch just like using a regular photo. This is extremely important for law enforcement application where a photo is often not available.

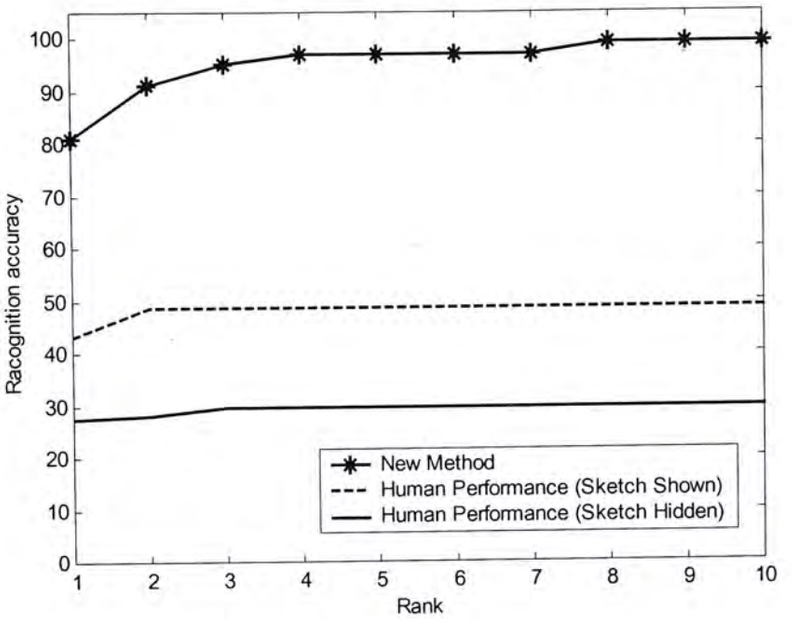


Figure 6-9 Comparison of accumulative match score between our automatic recognition method and human performance.

6.2 Face hallucination

In video surveillance, the faces of interest are often in small size because of the large distance between the camera and the objects. Image resolution becomes an important factor affecting face recognition performance. Since many detail facial features are lost in the low-resolution face images, the faces are often indiscernible. For identification, especially by human, it is useful to render a high-resolution face image from the low-resolution one. This technique is called face hallucination or face super-resolution [68][69].

The simplest way to increase image resolution is a direct interpolation of input images with such algorithms as nearest neighbour or cubic spline. However, the performance of direct interpolation is usually poor since no new information is added in the process. A number of super-resolution techniques have been proposed in recent years [6][7][15][30][31][43][44][64][67][68][72][82]. Most try to produce a super-resolution image from a sequence of low-resolution images [6][7][44][45][46][64][67] [72]. Some other approaches [30][31][82] are based on learning from training set containing high- and low- resolution image pairs, with the assumption that high-resolution images are Markov random field (MRF) [30][67][82]. These methods are more suitable for synthesizing local texture, and

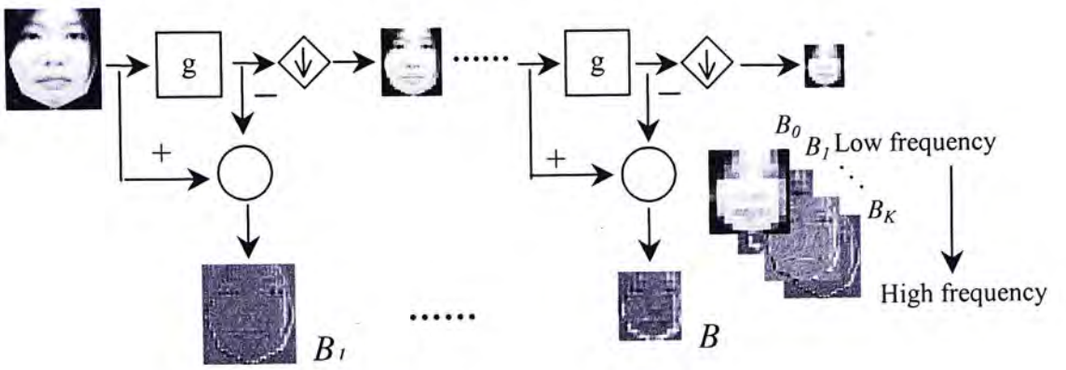


Figure 6-10 Multiresolution analysis in spatial domain.
g is the smoothing function, and $B_0, \ldots, B_K$ are different frequency bands

are usually applied to generic images without special consideration on the property of face images.

Baker and Kanade [68][69][70] develop a hallucination method based on the property of face image. Abandoning the MRF assumption, it infers the high frequency components from a parent structure by recognizing the local features from the training set. Liu et. al. [15] develop a two-step statistical modeling approach integrating global and local parameter models. Both of the two methods use complicated probabilistic models and are based on an explicit resolution reduction function, which is sometimes difficult to obtain in practice.

Since face images are well structured and have similar appearance, they span a small subset in the high dimensional image space [62][83]. In a study by Penev and Sirovich [59], face images are shown to be well reconstructed by PCA representation with 300-500 dimensions. Zhao et. al. [84] show that the dimensionality of face space is insensitive to image size. Moghaddam [8] down samples face images to 12 by 21 pixels and still achieves 95% recognition accuracy on 1800+ face images from the Feret database. These studies imply that facial components are highly correlated and the high frequency details of face images may be inferred from the low frequency components, utilizing the face structural similarity.

Resolution can be viewed as a kind of "style". Face images with different resolutions are in different spaces. Instead of using a probabilistic model, in [90] we apply eigentransformation to face hallucination. Using a small training set, the method can produce satisfactory results. Hallucination can effectively improve the resolution of face image, thus makes it much easier for a human being to

recognize a face. However, how much information has been extracted from the low-resolution image by the hallucination process and its contribution to automatic face recognition have not been studied before. In our method, PCA is applied to the low-resolution face image. In the PCA representation, different frequency components are independent. By selecting the number of eigenfaces, we could extract the maximum amount of facial information from the low-resolution face image and remove the noise. We also study the face recognition performance using different image resolutions. For automatic recognition, a low resolution bound is found through experiment. We find that hallucination may help the automatic recognition process, since it emphasizes the face difference by adding some high frequency details.

6.2.1 Multiresolution analysis

Viewing a 2D image as a vector, the process of getting a low-resolution face image from the high-resolution one can be formulated as

$$\vec{I}_l = H\vec{I}_h + \vec{n} . \tag{6-21}$$

Here, $\vec{I}_h$ is the high-resolution face image vector to be rendered, with length N as the total pixel number. $\vec{I}_l$ is the observed low-resolution face image vector with length $s^2 N$, where s is the downsampling factor ($0 < s < 1$). H is the transformation matrix involving blurring and downsampling process. The bluring operation can be estimated from the point-spread function of camera. In practice, it is often simplified as a Gaussian filter. The term $\vec{n}$ represents the noise perturbation to the low-resolution face image captured by camera. A detailed discussion on the super-resolution reconstruction constraints can be found in [68].

The hallucination problem can be discussed under the framework of multiresolution analysis. As shown in Figure 6-10, a process of iterative smoothing and downsampling decomposes the face image into different bands, $B_0, \ldots, B_K$. The low frequency component is encoded in the downsampled low-resolution image, and the difference between the original face image and the smoothed image contains the high frequency detail. In this decomposition, different frequency bands are not independent. Some components of the high-frequency bands, $B_1, \ldots, B_K$, can be inferred from the low frequency band B_0. This is a starting point for hallucination. Many super-resolution algorithms

assume the dependency as homogeneous Markov Random Fields (MRFs), i.e. the pixel only relies on the pixels in its neighborhood. This is an assumption for general images. It is not optimal for the face class without considering face structural similarity. A better way to address the dependency is using PCA, in which different frequency components are independent.

Many studies [52][59] on face space have shown that a face image can be reconstructed from eigenfaces in the PCA representation. Like the multiresolution analysis, PCA also decomposes face image into different frequency components. The difference is that the PCA method utilizes the face distribution to decorrelate face structure into independent frequency components, thus can encode face information more concisely. Our algorithm first employs PCA to extract as much useful information as possible from a low-resolution face images, and then renders a high-resolution face image by eigentransformation.

6.2.2 Eigentransformation for hallucination

We have a training set containing low-resolution face images, and corresponding high-resolution face images. Let $[\bar{l}_1,...\bar{l}_M]$ represent the low-resolution training face image set, from which the eigenfaces $E_l = [e_1,...,e_K]$ for low-resolution face images can be computed. As described in Section 6.1.1, apply PCA to the input low-resolution face image $\bar{x}_l$ to compute the principal components,

$$\bar{w}_l = E_l^T(\bar{x}_l - \bar{m}_l) \tag{6-22}$$

and $\bar{x}_l$ can be reconstructed as the linear combination of the low-resolution training face images,

$$\bar{r}_l = LV_l\Lambda_l^{-1/2}\bar{w}_l + \bar{m}_l = L\bar{c} + \bar{m}_l, \tag{6-23}$$

where $\bar{m}_l = \frac{1}{M}\sum_{i=1}^{M}\bar{l}_i$ is the mean face of the low-resolution training faces, $L = [\bar{l}'_1,...,\bar{l}'_M] = [\bar{l}_1 - \bar{m}_l,...,\bar{l}_M - \bar{m}_l]$, V_l and Λ_l are the eigenvectors matrix and eigenvalues matrix of L^TL, $E_l = LV_l\Lambda_l^{-1/2}$, and $\bar{c} = V_l\Lambda_l^{-1/2}\bar{w}_l = [c_1,...,c_M]^T$. Eq. (6-23) can be rewritten as,

$$\bar{r}_l = L\bar{c} + \bar{m}_l = \sum_{i=1}^{M} c_i\bar{l}'_i + \bar{m}_l \tag{6-24}$$

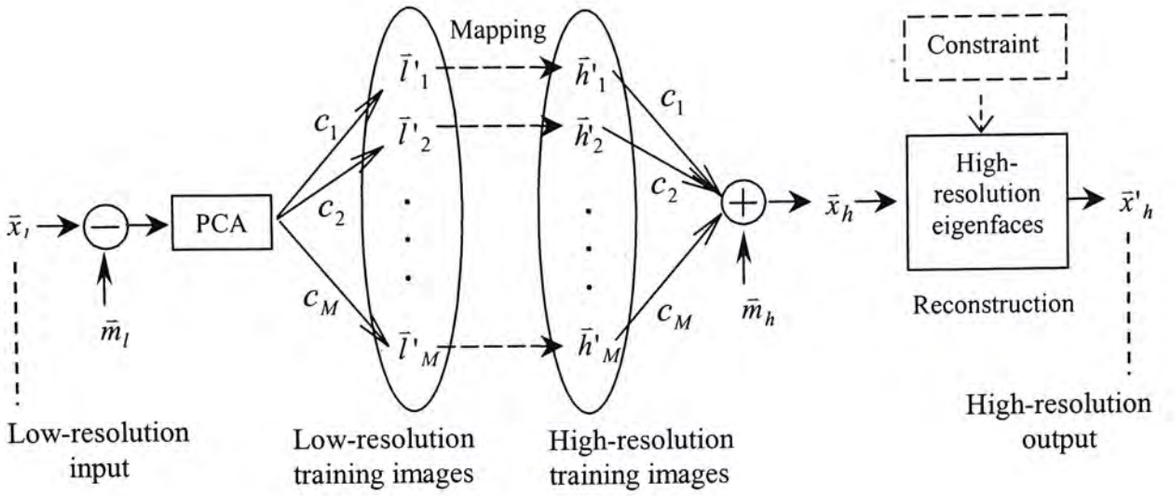


Figure 6-11 System diagram using eigentransformation for hallucination.

Here, $\bar{c}$ describes weight that each training face contributes in reconstructing the input face. The sample face that is more similar to the input face, has a greater weight contribution. Replacing each low-resolution image $\bar{l}'_i$ by its corresponding high-resolution sample $\bar{h}'_i$, and replacing $\bar{m}_l$ with the high-resolution mean face $\bar{m}_h$ we have,

$$\bar{x}_h = \sum_{i=1}^{M} c_i \bar{h}'_i + \bar{m}_h \,. \tag{6-25}$$

$\bar{x}_h$ is expected to be an approximation to the real high-resolution face image.

This reconstructed face should meet two necessary conditions in order to adequately approximate the original high-resolution face image. First, after resolution reduction of $\bar{x}_h$, the output should produce the low-resolution input face image. Second, $\bar{x}_h$ should be face-like at the high-resolution level. The first condition can be proved easily. From Eq. (6-21), without considering the noise perturbation, the transformation between high-resolution face image and low-resolution face image can be approximated as a linear operation. For the training set, we have

$$\bar{l}'_i = H\bar{h}'_i, \tag{6-26}$$

$$\bar{m}_l = H\bar{m}_h \,. \tag{6-27}$$

From (6-26) and (6-27), replacing $\bar{l}'$ and $\bar{m}_l$ with (6-26) and (6-27), we have

$$\bar{r}_l = \sum_{i=1}^{M} c_i H\bar{h}'_i + H\bar{m}_h = H\left(\sum_{i=1}^{M} c_i \bar{h}'_i + \bar{m}_h\right) = H\bar{x}_h \,. \tag{6-28}$$

For the second condition, Eq. (6-25) shows that $\vec{x}_h$ is the linear combination of high-resolution face images, so it should approximately be face-like at high-resolution level. Of course, some nonface-like distortion may be involved, since the combination coefficient c_i is not computed from the high resolution training data. We can reduce these nonface-like distortions by reconstructing $\vec{x}_h$ from the high-resolution eigenfaces. Let E_h and $\Lambda_h = diag(\lambda_1, \ldots, \lambda_K)$ be the eigenface and eigenvalue matrixes computed from the high-resolution training images. The principal components of $\vec{x}_h$ projecting on the high-resolution eigenfaces are

$$\vec{w}_h = E_h^T (\vec{x}_h - \vec{m}_h). \tag{6-29}$$

The eigenvalue λ_i is the variance of high-resolution face images on the ith eigenface. If the principal component $w_h(i)$ is much larger than λ_i, nonface-like distortion may be involved for the ith eigenface dimension. To reduce the distortion, we apply constraint on the principal component according using the eigenvalue,

$$\vec{w}'_h(i) = \begin{cases} w_h(i) & |w_h(i)| \le a\sqrt{\lambda_i} \\ sign(w_h(i)) * a\sqrt{\lambda_i} & |w_h(i)| > a\sqrt{\lambda_i} \end{cases}, \quad a > 0 \tag{6-30}$$

We use $a\sqrt{\lambda_i}$ to bound the principal components. Here, a is a positive scale parameter. The final hallucinated face image is reconstructed by

$$\vec{x}'_h = E_h^T \vec{w}'_h + \vec{m}_h. \tag{6-31}$$

The diagram of the hallucination algorithm based on eigentransformation is shown in Figure 6-11. When a low-resolution image $\vec{x}_l$ is input, it is approximated by a linear combination of the low-resolution images using the PCA method, and we get a set of coefficients $[c_1, c_2, \ldots, c_M]^T$ on the training set. Keeping the coefficients and replacing the low-resolution training images with the corresponding high-resolution ones, a new high-resolution face image can be synthesized. The synthesized face image is projected onto the high-resolution eigenfaces and reconstructed with constraints on the principal components. This transformation procedure is called eigentransformation, since it uses the eigenfaces to transform the input image to the output results.

6.2.3 Discussion

Similar to other multiscale analysis methods, PCA also decomposes face images into different frequency components. Figure 6-12 shows some eigenfaces, which are sorted by eigenvalues. Eigenfaces with large eigenvalues are "face-like", and characterize low frequency components. Eigenfaces with small eigenvalues are "noise-like", and characterize high frequency details. PCA is optimal for face representation because the K largest eigenfaces account for most of the energy and are most informative for face image set. The eigenface number K controls the detail level of the reconstructed face. As K increases, more details are added to the reconstructed face. Different from other multiscale analysis, in PCA, the frequency components are computed by decorrelation based on face structure, thus the different components in PCA are independent in probabilistic distribution. This property is important for the success of the eigentransformation algorithm.

In the eigentransformation algorithm, the hallucinated face image is synthesized by the linear combination of high-resolution training images and the combination coefficients come from the low-resolution face images using the PCA method. The algorithm improves the image resolution by inferring some high frequency face details from the low-frequency facial information by taking advantage of the correlation between the two parts. Because of the structural similarity among face images, in multiresolution analysis, there exists strong correlation between the high frequency band and low frequency band. For high-resolution face images, PCA can compact these correlated information onto a small number of principle components. Then, in the eigentransformation process, these principal components can be inferred from the principal components of the low-resolution face image by mapping between the high- and low-resolution training pairs. Therefore, some information in the high frequency bands are partially recovered.

In practice, the low-resolution image is often disturbed by noise that has a flat distribution on all the eigenvectors. For low-resolution face images, the energy on small eigenvectors is small, thus sometimes is overwhelmed by noise. The information on these noisy components (eigenfaces after K as shown in Fig. 6-13) is lost, and cannot be recovered since the components on different eigenvectors are independent in the PCA representation. By selecting an optimal eigenface number K, we can extract the facial information and remove the noise. Since $\bar{r}_l$ is

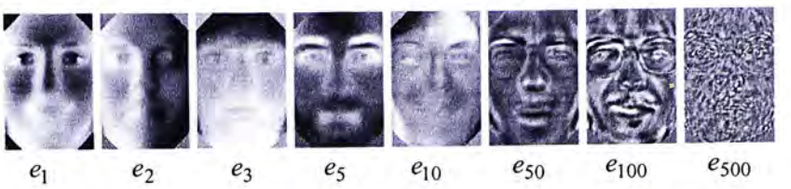


Figure 6-12 Eigenfaces sorted by eigenvalues. e_i is the ith eigenface.

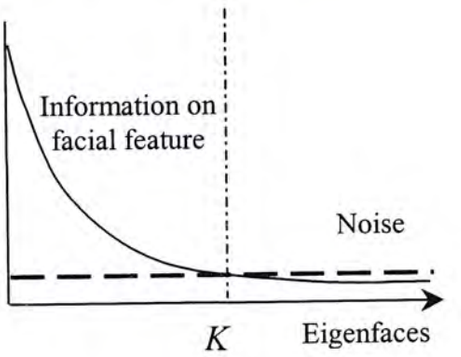


Figure 6-13 Extract facial information in the PCA space of low-resolution face image.

reconstructed from the K eigenfaces, given an optimal value of K, $\bar{r}_l$ encodes the maximum amount facial information recoverable in the low-resolution face image.

By adjusting the value of K, the eigentransformation method can control noise distortion. It makes full use of the facial information encoded in $\bar{r}_l$ to render high-resolution face image. We have shown that the hallucinated face image is face-like and could produce $\bar{r}_l$ after resolution reduction. Although these conditions do not guarantee that the hallucinated face image is exactly the same as the original high-resolution face image, it does provide a face-like possible solution to $\bar{r}_l$. This solution helps to infer high frequency components from the low frequency facial features, thus significantly improves the appearance of the face image.

We have noticed that some studies [8] use face images of small size for automatic face recognition, and have achieved satisfactory results. Through experiments, we would like to explore how the face resolution affects the recognition performance, and whether there is enough information for the low-resolution face images to distinguish different faces. Given the significant improvement of the face appearance by the hallucination process, it is interesting to investigate whether the hallucination helps automatic recognition. Since more high frequency details are recovered, we expect the hallucination process to help the recognition performance.

6.2.4 Experiment

6.2.4.1 Hallucination experiment

The hallucination experiment is conducted on a data set containing 188 individuals with one face image for each individual. Using the "leave-one-out" methodology, at each time, one image is selected for testing and the remaining are used for training. In preprocessing, the face images are aligned by the two eyes. The distance between the eye centers is fixed at 50 pixels, and the image size is fixed at 117×125. Images are blurred by averaging neighbour pixels and down sampled to low-resolution images. Here, we use the eye center distance *de* to measure the face resolution.

Some hallucination results are shown in Fig. 6-14. The input face images are down sampled to 23×25, with *de* equal to 10. Compared with the input image and the Cubic B-Spline interpolation result, the hallucinated face images have much clearer detail features. They are good approximation to the original high-resolution images.

We study the hallucination performance using different resolutions as input. The eye center distance is down sampled to 20, 15, 10, 7, and 5. An example is shown in Fig. 6-6, where (a) is the original face image. In Fig. 6-15 (b), the first row is input face images with different resolutions; the second row is the result of Cubic B-Spline interpolation; and the third row is the hallucination result. Figure 6-16 reports the average RMS error per pixel in intensity for the 188 face images under different resolutions. For a very low resolution, the low-resolution and direct interpolated face images are practically indiscernible, and the RMS error of Cubic B-spline interpolation increases quickly. The performance of hallucination by eigentransformation is much better. When *de* is down sampled to 10, the result of eigentransformation is still satisfactory. For further lower resolutions, there are some distortions on the eyes and mouth, but hallucinated images are still clear and face-like.

(a) input 23×25 (b) Cubic B-Spline (c) Hallucinated (d) Original 117×125

Figure 6-14 Hallucinated face images by eigentransformation.

In Fig. 6-17, we add zero mean Gaussian noise to the low-resolution face image. If no constraint is add to the principal components, the hallucinated face images in Fig. 6-17 (d) are with noise distortion and somewhat not face-like. Adding constraints on the principal components using Eq. (6-30), the reconstructed face images remove most of the noise distortion and retain most of the facial characteristics as shown in Fig. 6-17 (e).

As discussed in Section 6.2.3, some high frequency detail is lost in the process of bluring and downsampling, or overwhelmed by noise. Selecting the eigenface number in eigentransformation, we could control the detail level by keeping maximum facial information while removing most of the noise disturbance. In Fig. 6-18, we add zero mean Gaussian noises with four different standard deviations (σ) to the low-resolution face image with *de* equal to 10 (size of 23×25). The

image intensity is between 0 and 1. Two different eigenface number K, 50 and 180, are used in the eigentransformation. When only 50 eigenfaces are used in the eigentransformation, the hallucinated face images lose some individual characteristics. Although the edges and contours are clear, the hallucinated faces are more like a mean face. When eigenface number is equal to 180, more individual characteristics are added to the hallucinated face images. For relatively small noise ($\sigma = 0.03, 0.05$), these characteristics are similar to the original high-resolution face image. But for large noise ($\sigma = 0.1, 0.2$), even though the hallucinated faces are still face-like, the added characteristics start to deviate from those of true face. So when the noise is small, larger eigenface number is more suitable, since it can characterize the face better with more individual detail characteristics. When noise is large, small eigenvector number is better. Although the hallucinated faces contain less individual facial characteristics, it is more face-like. In practice, we could estimate the noise effect and choose the proper detail level for hallucination.

6.2.4.2 Recognition experiment

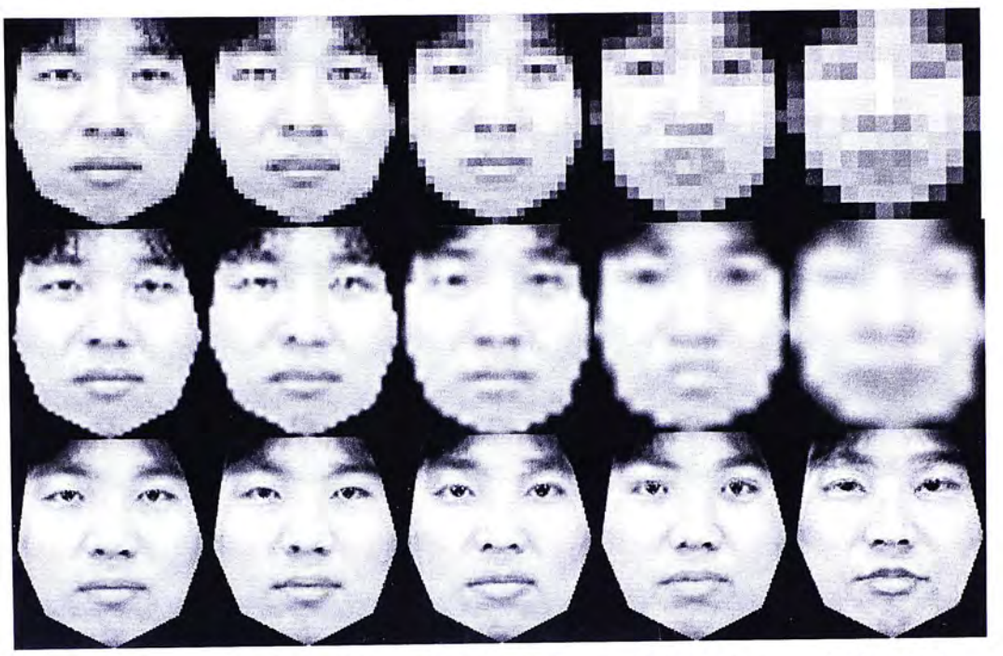
We study the recognition performance using low-resolution face images and hallucinated face images. Two hundred and ninety five individuals from the XM2VTS face database [37] are selected, with two face images in different sessions for each individual. One image is used as reference, and the other is used for testing. We use direct correlation for recognition, which is perhaps the simplest face recognition algorithm. The reason for using this simple classification algorithm is that our focus is on the comparison of recognition ability of the low-resolution and hallucinated face images rather than a sophisticated classification algorithm. The recognition accuracies over different resolutions are plotted in Fig. 6-19. When *de* is reduced from 50 to 10, there is only slight fluctuation on recognition accuracy using low-resolution face images. When *de* is further reduced to 7 and 5, the recognition accuracy for low-resolution face images drops greatly. Resolution with *de* equal to 10 is perhaps a lower bound for recognition. Below this level there may not be enough information for recognition. This is also consistent with the hallucination experiment in Fig. 6-15. Satisfactory hallucination results can be obtained when *de* is equal to or larger than 10.

We also try to explore whether hallucination can contribute to automatic face recognition. We expect hallucination to make the recognition procedure easier, since it emphasizes the face difference by adding some high frequency details. In this experiment, the low-resolution testing image is hallucinated by reference face images, but the face image of the testing individual is excluded from the training set. As shown in Fig. 6-19, the hallucination improves the recognition accuracy when the input face images have very low resolutions. The improvement seems not as significant as the improvement in the face appearance. Further investigation in psychology study may be needed to address this phenomenon. It seems that human visual system can better interpret the added high frequency details.

6.2.4.3 Conclusion

Because of the structural similarity, face images can be synthesized from the linear combination of other samples. Based on this property of face images, hallucination can be implemented by eigentransformation. By selecting the frequency level in the PCA representation, our method extracts maximum facial information from the low-resolution face images and is robust to noise. The resolution and quality of face images are greatly improved over the low-resolution images. The hallucination process not only helps a human being to identify faces but also makes the automatic face recognition procedure easier. It will be interesting to study why the hallucinated image is significantly better perceived by a human being than by the automatic recognition system.

(a) Original 50 (117×125)



20 (47×50) 15 (35×37) 10 (23×25) 7 (16×17) 5 (11×12)

(b) The first row is the input face images, whose eye center distances have been down sampled to 20, 15, 10, 7, and 5 pixels respectively; the second row is the interpolation result using Cubic B-Spline; the third row is the hallucinated face images.

Figure 6-15 Hallucinated face images using different resolutions as the input.

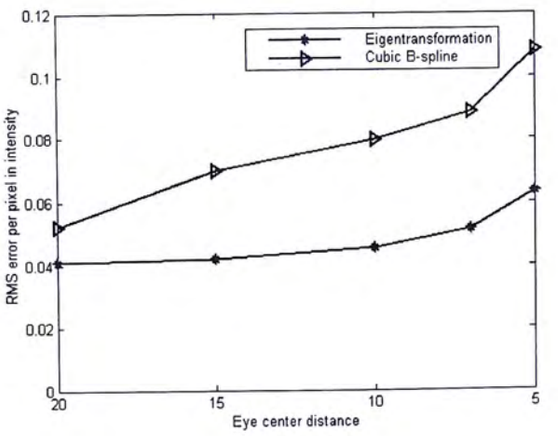


Figure 6-16 RMS error per pixel in intensity using Cubic spline interpolation and hallucination by eigentransformation.

The intensity is between 0 and 1. The resolution is marked by eye center distance

(a) (b) (c) (d) (e)

Figure 6-17 Adding constrain to the principal components of the hallucinated face images.

(a): Original high resolution face images; (b): Low-resolution face images with *de*=10, and the size 23×25; (c): Low resolution face images added zero mean 0.03 standard variation Gaussian noise; (d): Hallucinated face images from (c) without constraints on the principal components; (e): Hallucinated face images from (c) with constraints on the principal components. (e) is more face-like and less noisy comparing to (d), and it retains most of the facial characteristics of (d).

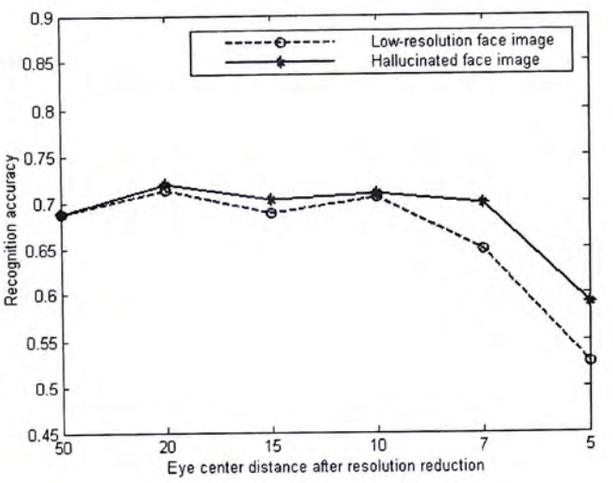
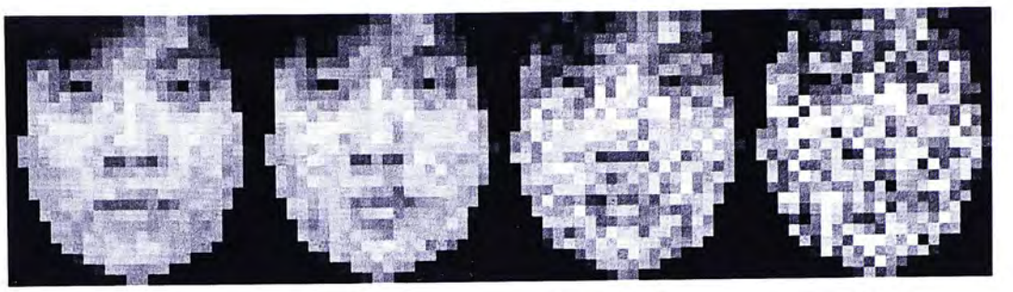


Figure 6-18 Recognition accuracy using low-resolution face images and hallucinated face images based on XM2VTS database.

The resolution is marked by eye center distance with 50, 20, 15, 10, 7, and 5.

(a)



($\sigma = 0.03$) ($\sigma = 0.05$) ($\sigma = 0.1$) ($\sigma = 0.2$)

(b)

(K=50, $\sigma = 0.03$) (K=50, $\sigma = 0.05$) (K=50, $\sigma = 0.1$) (K=50, $\sigma = 0.2$)

(K=180, $\sigma = 0.03$) (K=180, $\sigma = 0.05$) (K=180, $\sigma = 0.1$) (K=180, $\sigma = 0.2$)

(c)

Figure 6-19 Hallucinating face with additive zero mean, Gaussian noise.

The input face image is 23×25. (a): Original high-resolution face image; (b): Low-resolution face images with noise; (c): Hallucinated face images using different eigenface number. K is the eigenface number in eigentransformation, and σ is the standard variation of Gaussian noise.

6.3 Discussion

In this section, we study the matching between face images with significant and definite transformation, which is referred as "style". The transformation is so significant that the face image cannot be directly matched using standard subspace

methods. However, for a kind of style, the transformation is not arbitrary, but can be approximated as some definite function. For a face image for matching, the style is previous known, and some training face images undergoing the same transformation can be collected to learn the transformation procedure. In the study of this section, photo-sketch, and face images with different resolutions are all viewed as different stylistic face images. For more examples, face images wearing glasses, with some fixed poses and under fixed lighting source all can be viewed as undergoing some definite transformations. To eliminate the transformation difference, a normal way is to derive the explicit transformation function, such as using 3D model to normalize pose and lighting changes. However, in many cases, the transformation function is too complex to be computed. The advantage of eigentransformation is that it is able to perform the transformation using the one-to-one mapping between training samples without knowing the transformation function. It utilizes the face structural similarity. Face images with the same style construct a subspace. Eigentransformation realizes the transformation between different subspaces. Since face images are in different modalities, gray levels in pixels or eigenfaces are not identical in different subspaces. Eigentransformation assumes that the weights $\vec{c}$ on training samples are somewhat invariant to transformations. In intuition, the more similar is the training sample to the input face, the greater is its weight on reconstruction. We have proved that this assumption is reasonable when transformation can be approximated as linear. In this thesis, eigentransformation is applied to face sketch recognition and hallucination. However, we have shown that it also can be applied to normalizing pose transformation and remove the effect of glass. This is a direction for our further study.

Chapter 7
Conclusion

In this thesis, we extensively study the subspace methods for face recognition. The unique contribution, disadvantage, and relationship of different subspace methods can be well understood under the proposed framework. The framework breaks down the limits on conventional subspace methods. It unifies and improves different subspace methods to attain the best recognition performance using a unified subspace analysis. It is helpful to develop novel subspace methods for face recognition. Here, we would like to discuss several possible directions for further study.

In this thesis, all the discussion for subspace analysis is based on linear projection. However, in practice the face distribution can be more complex and a linear PCA on the face difference set may not be able to decorrelate higher order dependencies of different components, thus $\tilde{I}$, $\tilde{T}$, and $\tilde{N}$ may not be fully separated. A potential solution to this problem is to project the image vectors into a higher dimensional space by a nonlinear function. Based on this consideration, kernel PCA and kernel LDA have been developed recently. With the new framework, the starting point of the kernel approaches does not have to be limited to the standard subspace methods. We can propose a kernel based improvement based on the best point in the new parameter space.

When there are too many faces in the dataset, the face distribution may be too complex to be classified using a linear subspace. Some approaches, such as LDA mixture model [25], have been proposed to partition the face classes in the gallery into several clusters, and compute the discriminnat vectors for each cluster respectively. The classification problem seems to be simplified, since the class number in each cluster is reduced. However, training sample number in each cluster is also reduced, so the transformation difference cannot be well estimated. Furthermore, in this approach, the training samples must come from the face classes in the gallery. When the face class in the gallery has few samples, it will heavily suffer from the small sample size problem. Our framework first proposes

to use different training data in different steps. It can effectively release the difficulty encountered by the mixture model.

In this thesis, we apply eigentransformation to face sketch recognition and hallucination. It also can be used to remove the great transformation difference caused by poses changes or wearing glasses, since face images wearing glasses or under fixed posed all can be viewed as undergoing some definite transformations. The subspace face recognition system can be further improved. When a probe face image is input, first judge whether it undergoes some definite transformation, and use eigentransformation to eliminate the great transformation difference. Then a unified subspace analysis is applied to extract the discriminant features for recognition.

Publication List of This Thesis

[1] X. Wang, and X. Tang, "Unified Subspace Analysis for Face Recognition," *Proceedings of IEEE International Conference on Computer Vision*, 2003.

[2] X. Tang, and X. Wang, "Face Sketch Synthesis and Recognition," *Proceedings of IEEE International Conference on Computer Vision*, 2003.

[3] X. Wang, and X. Tang, "An improved Bayesian Algorithm in Reduced PCA Space," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. III_129-III_132, 2003.

[4] X. Tang, and X. Wang, "Face Photo Recognition Using Sketch," *Proceedings of IEEE International Conference on Image Processing*, pp. I_257-I_260, 2002.

[5] X. Wang, and X. Tang, "Face Hallucination and Recognition," *Proceedings of the 4th International Conference on Audio- and Video-Based Person Authentication,* 2003.

[6] X. Wang, and X. Tang, "A Unified Framework for Subspace Face Recognition," Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, under the second round review.

[7] X. Tang, and X. Wang, "Face Sketch Recognition," Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, under the second round review.

[8] X. Wang, and X. Tang, "Hallucinating Face by Eigentransformation," Submitted to IEEE Special Issue on Biometric Systems of IEEE Trans. on Systems, Man, and Cybernetics, Part C.

[9] X. Wang, and X. Tang, "Dual-Subspace LDA for High Dimensional Data Classification," Submitted to IEEE Transactions on Image Processing.

Bibliography

[1] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic Interpretation and Coding of Face Images Using Flexible Models," *IEEE Trans. on PAMI,* Vol. 19, No. 7, pp. 743-756, July 1997.

[2] A. Levy, and M. Lindenbaum, "Sequential Karhunen-Loeve Basis Extraction and its Application to images," *IEEE Trans.* on Image Processing, Vol. 9, No. 8, pp. 1371-1374, August. 2002.

[3] A. M. Martinez, and A. C. Kark, "PCA versus LDA," *IEEE Trans. on PAMI,* Vol. 23, No. 2, pp. 228-233, Feb. 2001.

[4] A. M. Martinez, and R. Benavente, "The AR Face Database," *CVC Technical Report #24,* June 1998.

[5] A. Pentland, B. Moghaddam, T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proceedings of CVPR,* pp. 84-91, 1994.

[6] A. Patti, M. Sezan, and A. Tekalp, "Super-resolution Video Reconstruction with Arbitrary Sampling Latices and Nonzero Aperture Time," *IEEE Trans. on Image Processing*, Vol. 6, No. 8, pp. 1064-1076, 1997.

[7] B. Bascle, A. Blake, and A. Zisserman, "Motion Deblurring and Super-Resolution from an Image Sequence," *Proceedings of ECCV*, pp. 573-581, 1996.

[8] B. Moghaddam, "Principle manifolds and probabilistic subspace for visual recognition," *IEEE Trans. on PAMI*, Vol. 24, No. 6, pp. 780-788, June, 2002.

[9] B. Moghaddam, and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. on PAMI*, Vol. 19, No. 7, pp. 775-779, July, 1997.

[10] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian Face Recognition," *Pattern Recognition,* Vol. 33, pp. 1771-1782, 2000.

[11] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond Eigenface: Probabilistic Matching for Face Recognition," *Proceedings of International Conference on Automatic Face and Gesture Recognition (FG'98),* pp. 30-35, April, 1998.

[12] B. J. Frey, A. Colmenarez, and T. S. Huang, "Mixtures of Local Linear Subspaces for Face Recognition," *Proceedings of CVPR,* pp. 32-37, 1998.

[13] B. Scholkopf, A. Smola, K. Muller, "Kernel Principal Component Analysis," In B. Scholkopf, C. Burges, and A. Smola, editors, Advances in Kernel Method-Support Vector Learning, pp. 327-352, MIT Press, 1999.

[14] C. Ki-Chung, K. S. Cheol, K. S. Ryong, "Face Recognition Using Principal Component Analysis of Gabor Filter Responses," *Proceedings of International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems,* pp. 53-57, 1999.

[15] C. Liu, H. Shum, and C. Zhang, " A Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,* pp. 192-198, 2001.

[16] C. Liu and H. Wechsler, "Enhanced Fisher Linear Discriminant Models for Face Recognition," *Proceedings of ICPR,* Vol. 2, pp. 1368-1372, 1998.

[17] C. Liu, and H. Wechsler, "Evolutionary Persuit and its Application to Face Recognition," *IEEE Trans. on PAMI*, Vol. 22, No. 6, pp. 570-582, June, 2000.

[18] C. Liu, and H. Wechsler, "A Shape- and Texture-Based Enhanced Fisher Classifier for Face Recognition," *IEEE Trans. on Image Processing,* Vol. 10, No. 4, pp. 598-608, April, 2001.

[19] C. Liu, and H. Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition," *IEEE Trans. on Image Processing,* Vol. 11, No. 4, pp. 467-476, April, 2002.

[20] C. L. Kotropoulos, A. Tefas, and I. Pitas, "Frontal Face Authentication Using Discriminating Grids with Morphological Feature Vectors," *IEEE Trans. on MultiMedia,* Vol. 2, pp. 14-26, March 2000.

[21] D. L. Swets and J. Weng, "Discriminant Analysis and Eigenspace Partition Tree for Face and Object Recognition from Views," *Proceedings of International Conference on Automatic Face Gesture Recognition,* pp. 192-197, 1996.

[22] D. Swets, J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. on PAMI*, Vol. 16, No. 8, pp. 831-836, August, 1996.

[23] G. Baudat, and F. Anouar, "Generalized Discriminant Analysis Using a Kernel Approach," Neural Computation, 12 (10): 2385-2404, 2000.

[24] G. W. Stewart, "Introduction to Matrix Computations," Academic Press, New York, 1973.
[25] H. Kim, D. Kim, and S. Y. Bang, "Face Recognition Using LDA Mixture Model," *Proceedings of ICPR*, pp. 486-489 2002.
[26] H. Koshimizu, M. Tominaga, T. Fujiwara, and K. Murakami. "On Kansei Facial Processing for Computerized Facial Caricaturing System PICASSO," *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics,* Vol. 6, pp. 294-299, 1999.
[27] H. M. El-Bakry, and M. A. Abo-Elsoud, "Integrating Fourier Descriptors and PCA with Neural Networks for Face Recognition," *Seventh National Radio Science Conference,* Feb. 2000.
[28] H. Yu, and J. Yang, "A Direct LDA Algorithm for High-Dimensional Data-with Application to Face Recognition," *Pattern Recognition,* Vol. 34, pp. 2067-2070, 2001.
[29] I. Craw, N. Costen, T. Kato, and S. Akamatsu, "How Should We Represent Faces for Automatic Recognition?" *IEEE Trans. on PAMI,* Vol. 21, No.8, pp. 725-736, August 1999.
[30] J.D. Bonet, "Multiresolution sampling procedure for analysis and synthesis of texture images," *Proceedings of SIGGRAPH 97*, pp. 361-368, 1997.
[31] J. De Bonet and P. Viola, "A non-parametric multi-scale statistical model for neutral images," *Advances in Neutral Information Processing*, 10, 1997.
[32] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas, "On Combining Classifiers," IEEE Trans. on PAMI, Vol. 20, N0. 3, pp. 226-239, March 1998.
[33] J. Zhang, Y. Yan, and M. Lades, "Face Recognition: Eigenface, Elastic Matching, and Neural Net," *Proceedings of IEEE,* Vol. 85, No. 9, pp.1423-1435, Sep. 1997.
[34] K. Chang and J. Ghosh, "A Unified Model for Probabilistic Principal Surfaces," *IEEE Trans. on PAMI,* Vol. 23, No. 1, pp. 22-41, Jan. 2001.
[35] K. Etemad, and R. Chellappa, "Discriminant Analysis for Recognition of Human Faces," *Proceedings of International Conference on Acoustics, Speech and Signal Processing,* pp. 2148-2151, 1996.
[36] K. Fukunnaga, "Introduction to Statistical Pattern Recognition," Academic Press, second edition, 1991.
[37] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," *Proceedings of International Conference on Audio- and Video-Based Person Authentication*, pp. 72-77, 1999.
[38] L. Chen, H. Liao, M. Ko, J. Liin, and G. Yu, "A New LDA-based Face Recognition System Which can Solve the Samll Sample Size Problem," *Pattern Recognition*, Vol. 33, No. 10, pp. 1713-1726, Oct. 2000.
[39] L. Xu, A. Krzyzak, C. Y. Suen, "Method of Combining Multiple Classifiers and Their Applications to Handwriting Recognition," IEEE Trans. on System, Man, and Cybernetics, Vol. 22, No. 3, 418-435, 1992.
[40] L. Wiskott, J.M. Fellous, N. Kruger and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 19, No.7, pp. 775-779, July, 1997.
[41] M. A. Kramer, "Nonlinear Principle Components Analysis Using Autoassociative Neural Networks," *Am. Inst. Chemical Eng. J.,* Vol. 32, No. 2, pp. 1010, 1991.
[42] M. Bichsel, and A. Pentland, "Human face recognition and the face image set's topology," *CVGIP: Image Understanding*, Vol. 59, pp. 254-261, 1994.
[43] M. Chiang, and T.E. Boult, "Local Blur Estimation and Super-Resolution," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,* pp. 821-826, 1997.
[44] M. Elad and A. Feuer, "Restoration of A Single Superresolution Image from Several Blurred Noisy and Undersampled Measured Images," *IEEE Trans. on Image Processing,* Vol. 6, No. 12, pp. 1646-1658, 1997.
[45] M. Elad and A. Feuer, "Super-Resolution Reconstruction of Image Sequences," *IEEE Trans. on PAMI,* Vol. 21, No. 9, 1999.
[46] M. Elad and A. Feuer, "Super-Resolution of an Image Sequence-Adaptive Filtering Approach," *IEEE Trans. on Image Processing*, Vol. 8, No. 3, pp. 387-395, 1999.
[47] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for Characterization of Human Faces," *IEEE Trans. on PAMI,* Vol. 12, No.1, pp. 103-108, Jan. 1990.
[48] M. S. Bartlett, T.J. Sejnowski, "Independent of Face Images: A Representation for Face Recognition," *Proceedings of the Fourth Annual Joint Symposium on Neural Computation,* Pasadena, CA, May 17, 1997.

[49] M. Skurichina, and R. P. W. Duin, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers," Pattern Analysis & applications, pp. 121-135, 2002.
[50] M. Turk, "A Random Walk through Eigenspace," *IEICE Trans. Information and Systems*, Vol. E84-D, No. 12, pp. 1586-1595, Dec. 2001.
[51] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71-86, 1991.
[52] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proceedings of CVPR,* pp. 586-591, Hawaii, June, 1991.
[53] M. Yang, N. Ahuja and D. Kriegman, "Face Recognition Using Kernel Eigenfaces," *Proceedings of IEEE, ICIP*, Vol. 1, pp. 37-40, 2000.
[54] P. Comon, "Independent Component Analysis-A New Concept?" *Signal Processing*, Vol. 36, pp. 287-314, 1994.
[55] P. C. Yunen, and J.H. Lai, "Face Representation Using Independent Component Analysis," *Pattern Recognition,* Vol. 35, pp. 1247-1257, 2002.
[56] P. J. Phillips, H. Moon, and S. A. Rozvi, "The Feret Evaluation Methodolody for Face Recognition Algorithms," *IEEE Trans. PAMI*, Vol. 22, No. 10, pp. 1090-1104, Oct. 2000.
[57] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET Evaluation," in *Face Recognitioin: From Theory to Applications,* H. Wechsler, P. J. Phillips, V. Bruce, F.F. Soulie, and T.S. Huang, Eds., Berlin: Springer-Verlag, 1998.
[58] P.N. Belhumeur, J. Hespanda, and D. Kiregeman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. on PAMI,* Vol. 19, No. 7, pp. 711-720, July 1997.
[59] P. S. Penev and L. Sirovich, "The Global Dimensionality of Face Space", *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition,* pp. 264-270, 2000.
[60] Q. Liu, R. Huang, H. Lu and S. Ma, "Kernel-Based Optimized Feature Vectors Selection and Discriminant Analysis for Face Recognition," *Proceedings of ICPR*, pp. 362-365, 2002.
[61] R. Brunelli and T. Poggio, "Face Recognition: Features Versus Templates," *IEEE Trans. on PAMI,* Vol. 15, No. 10, pp. 1042-1052, Oct. 1993.
[62] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces: A Survey," *Proceedings of the IEEE*, Vol. 83, pp. 705-741, May 1995.
[63] R.G. Uhl and N.d.V. Lobo, "A Framework for Recognizing a Facial Image from A Police Sketch," *Proceedings of CVPR*, pp. 586-593, 1996.
[64] R. Hardie, K. Barnard, and E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. on Image Processing*, Vol. 6, No. 12, pp. 1621-1633, 1997.
[65] R. J. Baron, "Mechanisms of Human Facial Recognition," *Int. J. Man Machine Studies,* Vol. 15. 137-178, 1981.
[66] R.P.W. Duin, and R. Haeb-Umbach, "Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria," *IEEE Trans. on PAMI,* Vol. 23, No. 7, pp. 762-766, July, 2001.
[67] R. Schultz, and R. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. on Image Processing,* Vol. 5, No. 6, pp. 996-1011, 1996.
[68] S. Baker, and T. Kanade, "Limits on Super-Resolution and How to Break them," *IEEE Trans. on PAMI*, Vol. 24, No. 9, pp. 1167-1183, 2002.
[69] S. Baker, and T. Kanade, "Hallucinating Faces," *Proceedings IEEE International Conference on Automatic Face and Gesture Recognition,* pp. 83-88, 2000.
[70] S. Baker, and T. Kanade, " Limits on Super-Resolution and How to Break them," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2000.
[71] S. B. Yacoub, Y. Abdeljaoud, and E. Mayoraz, "Fusion of Face and Speech Data for Person Identity Verification," *IEEE Transactions on Neural Networks*, pp. 1065-1074, Vol. 10, No. 5, September 1999.
[72] S. Kim, and W. Su, "Recursive Reconstruction of High-Resolution Reconstruction of Blurred Multiframes Images," *IEEE Trans. on Image Processing*, Vol. 2, pp. 534-539, 1993.
[73] T. Cooke, "Two Variations on Fisher's Linear Discriminant for Patter Recognition," *IEEE Trans. on PAMI,* Vol. 24, No. 2, pp. 268-273, Feb. 2002.
[74] T. F. Cootes and C. J. Taylor, "Statistical Model of Appearance for Computer Vision," *Technical Report,* University of Manchester, Manchester M13 9PT, U. K., 2001.
[75] T. F. Cootes and C. J. Edwards, and C. J. Taylor, "Active Appearance Models," IEEE Trans. on PAMI, Vol. 23, No. 6, pp. 681-685, June, 2001.

[76] T. Kam Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. on PAMI*, Vol. 20, No. 8, pp. 832-844, August 1998.
[77] T. Kam Ho, "Nearest Neighbour in Random Subspace," *Intelligent Data Analysis*, 3, pp. 191-209, 1999.
[78] T. Kam Ho, J. Hull, S. Srihari, "Decision Combination in Multiple Classifier Systems," IEEE Trans. on PAMI, Vol. 16, No.1, pp. 66-75, Jan. 1994.
[79] T. Shakuaga, and K. Shigenari, "Decomposed Eigenface for Face Recognition Under Various Lighting Conditions," *Proceedings of CVPR,* pp. 864-871, 2001.
[80] W. Hwang, J. Weng, "Hierarchical Discriminant Regression," *IEEE Trans. on PAMI,* Vol. 22, No. 11, pp. 1277-1293, Nov. 2000.
[81] W. Konen, "Comparing Facial Line Drawings with Gray-Level Images: A Case Study on PHANTOMAS," *Proceedings of International Conference on Artifical Neural Networks,* pp. 727-734, 1996.
[82] W.T. Freeman and E.C. Pasztor, "Learning Low-Level Vision," *Proceedings of IEEE International Conference on Computer Vision*, 1999.
[83] W. Zhao, R. Chellappa, and P. Phillips. "Face Recognition: A Literature Survey," *Technical Report*, 2002.
[84] W. Zhao, R. Chellapa, and P. Philips, "Subspace Linear Discriminant Analysis for Face Recognition," *Technical Report CAR-TR-914,* 1996.
[85] W. Zhao, R. Chellappa, and N. Nandhakumar, "Empirical Performance Analysis of Linear Discriminant Classifiers," *Proceedings of CVPR,* pp. 164-169, 1998.
[86] X. Tang, and X. Wang, "Face Photo Recognition Using Sketch," Proceedings of ICIP, pp. I-257-I-260, 2002.
[87] X. Wang, and X. Tang, "Unified Subspace Analysis for Face Recognition," *Proceedings of ICCV*, 2003.
[88] X. Tang, and X. Wang, "Face Sketch Synthesis and Recognition," *Proceedings of ICCV*, 2003.
[89] X. Wang, and X. Tang, "An improved Bayesian Algorithm in Reduced PCA Space," *Proceedings of ICASSP*, 2003.
[90] X. Wang, and X. Tang, "Face Hallucination and Recognition," *Proceedings of the 4th International Conference on Audio- and Video-Based Person Authentication,* 2003.
[91] Y. Moses, Y. Adini, and S. Ullman, "Face Recognition: the Problem of Compensating for Changes in Illumination Direction," Proceedings of Third European Conference on Computer Vision, pp. 286-295, 1994.

CUHK Libraries
004076634