NEW SUBSPACE APPROACHES IN PATTERN RECOGNITION

Mehmet KOÇ

Ph.D. Dissertation

Graduate School of Sciences

Electrical and Electronics Engineering

December 2012

ANADOLU ÜNİVERSİTESİ

JÜRİ VE ENSTİTÜ ONAYI

Mehmet Koç'un **"Örüntü Tanımada Yeni Altuzay Yaklaşımları"** başlıklı **Elektrik-Elektronik Mühendisliği** Anabilim Dalındaki Doktora Tezi 18.10.2012 tarihinde, aşağıdaki jüri tarafından Anadolu Üniversitesi Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

	Adı Soyadı	İmza
Üye (Tez Danışmanı)	**: Prof. Dr. ATALAY BARKANA**	**.....**
Üye	**: Prof. Dr. ÖMER NEZİH GEREK**	**.....**
Üye	**: Prof. Dr. VAKIF CAFER**	**.....**
Üye	**: Prof. Dr. REFAİL KASIMBEYLİ**	**.....**
Üye	**: Doç. Dr. RİFAT EDİZKAN**	**.....**

Anadolu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ..... tarih ve ..... sayılı kararıyla onaylanmıştır.

Enstitü Müdürü

ABSTRACT

Ph.D. Dissertation

NEW SUBSPACE APPROACHES IN PATTERN RECOGNITION

Mehmet KOÇ

Anadolu University

Graduate School of Sciences

Electrical and Electronics Engineering Program

Supervisor: Prof. Dr. Atalay BARKANA

2012, 80 pages

In this thesis, three topics of pattern recognition namely, feature selection, single image per subject problem, and within-class scatter matrix estimation of classes are discussed. One of the important factors that affect the performance of the pattern recognition system is the dimension of the feature vector. A novel feature selection method is proposed which is related to DCVA in order to overcome the high dimensionality problem encountered in recognition issues. The important features are determined by the column norms of the projection matrix to the range space of all common vectors. Another factor that affects the performance of the pattern recognition system is the training sample size. Traditional methods which use the within-class scatter matrix fail if one sample from each subject is available because the within-class scatter matrices are all zero. An image decomposition method that uses QR-decomposition with column pivoting (QRCP) is proposed to overcome one sample per class problem. Also a two-dimensional extension of DCVA is proposed. The third and also important problem is to make a good estimation of within-class scatter matrices. But in high dimensional classification problems generally it is not possible to find a sufficient number of samples per class. In our proposal, at first the data is projected onto the range space of the total within-class scatter matrix. Then the within-class scatter matrix of a class is modeled using not only its own data but also the data of all other classes.

Keywords: Face recognition; Subspace methods; Feature selection; One sample problem; Discriminative common vector; Within-class scatter matrix

ÖZET

Doktora Tezi

ÖRÜNTÜ TANIMADA YENİ ALTUZAY YAKLAŞIMLARI

Mehmet KOÇ

Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik-Elektronik Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Atalay BARKANA

2012, 80 sayfa

Bu tezde, örüntü tanımadaki üç konu olan öznitelik seçimi, her sınıftan tek örnek problem ve sınıfların sınıf-içi saçılım matrisi tahmini üzerine çalışılmıştır. Örüntü tanıma sisteminin başarımına etki eden önemli etkenlerden biri öznitelik vektörü boyutudur. Tanıma problemlerinde karşılaşılan yüksek boyut sorununun üstesinden gelmek için AOVY ile ilişkili bir öznitelik seçimi yöntemi önerilmiştir. Özniteliklerin önemli olanları tüm ortak vektörlerin görüntü uzayına izdüşüm matrisinin sütun normları ile belirlenir. Örüntü tanıma sistemlerini etkileyen diğer bir etken de eğitim örneklem büyüklüğüdür. Sınıf-içi saçılım matrisini kullanan klasik yöntemler her sınıftan bir örnek varsa başarısız olmaktadırlar, çünkü sınıf-içi saçılım matrislerinin hepsi sıfır olmaktadır. Tek örnek probleminin üstesinden gelmek için pivot yöntemi ile QR ayrıştırmasını (QRCP) kullanan bir resim ayrıştırma yöntemi önerilmiştir. Ayrıca AOVY yönteminin iki boyutlu bir genişletmesi önerilmiştir. Üçüncü önemli problem ise bir sınıfın sınıf-içi saçılım matrisinin tahmininin iyi yapılabilmesidir. Fakat yüksek boyutlu sınıflandırma problemlerinde her sınıf için yeterli sayıda örnek bulmak genellikle mümkün değildir. Bizim önerimizde, ilk olarak verilerin toplam sınıf-içi saçılım matrisinin görüntü uzayına izdüşümü alınır. Daha sonra bir sınıfın sınıf-içi saçılım matrisi sadece kendi verisi kullanılarak değil diğer sınıfların verileri de kullanılarak modellenir.

Anahtar Kelimeler: Yüz tanıma; Altuzay yöntemleri; Öznitelik seçimi; Tek örnek problemi; Ayırt edici ortak vektör; Sınıf-içi saçılım matrisi

ACKNOWLEDGEMENT

I would like to thank my advisor, Prof. Dr. Atalay BARKANA for his support and guidance throughout my thesis work. It would be very hard to complete this work without his advisory. I would like to thank Prof. Dr. Ömer Nezih GEREK and Prof. Dr. Vakıf CAFER for their constructive comments.

I would also thank to my wife Arife for her patience and moral support throughout this exhausting work. I want to express my gratitude to my father for his confidence and support.

Mehmet KOÇ

December, 2012

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

$a, b, c, ...$	: Scalars
$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, ...$	: Vectors
$\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, ...$	: Matrices
I	: Identity matrix
FLDA	: Fisher Linear Discriminant Analysis
2D-FLDA	: Two Dimensional Fisher Linear Discriminant Analysis
SVD	: Singular Value Decomposition
CVA	: Common Vector Approach
LRC	: Linear Regression Classification
PCA	: Principal Component Analysis
1D-DCVA	: Discriminative Common Vector Approach
2D-DCVA	: Two Dimensional Discriminative Common Vector Approach
QRCP	: QR Decomposition with Column Pivoting
SVM	: Support Vector Machines
B_j	: Difference subspace of j^{th} class
$B_j^{\perp}$	: Indifference subspace of j^{th} class
$\boldsymbol{a}_{com}^{j}$	: Common vector of j^{th} class
$\boldsymbol{A}_{com}^{j}$	: Common matrix of j^{th} class
$\boldsymbol{\Phi}_j$	: Within-class covariance matrix of j^{th} class
$\boldsymbol{\Phi}_T$	: Total scatter matrix
$\boldsymbol{\Phi}_{com}$	: Covariance matrix of common vectors
$\boldsymbol{P}$	: Projection matrix onto the difference subspace

$\boldsymbol{P}^{\perp}$	: Projection matrix onto the indifference subspace
$\boldsymbol{S}_W$	: Within-class covariance matrix
$\boldsymbol{S}_B$	: Between-class covariance matrix
$\boldsymbol{A}_j^i$	: j^{th} image from i^{th} class
$\|\cdot\|_2$	: Euclidean distance
$\mathbb{R}^{m\times n}$	: m by n dimensional real space
σ_i	: i^{th} singular value
$R(\boldsymbol{\Phi}_T)$	: Range space of $\boldsymbol{\Phi}_T$
SES	: Sum of error squares
∇f	: Gradient of function f

1 INTRODUCTION

Face recognition is an active research area for pattern recognition and computer vision and it is a difficult and complex problem. A face recognition system is for automatically identifying or verifying a person from his/her digital image. Face recognition has many application areas such as security, person identification, passports, information security, law enforcements [1,2].

There are several parameters that affect the performance of the face recognition system. One of the most important is the dimension of the feature vectors. The digital images contain large number of pixels which are represented with gray level values. Each image corresponds to a point in high dimensional space and this increases the computational and storage costs of the face recognition system [3,4,5]. Mutually related features in a feature vector may result in small recognition gains but this increases the computational cost even if they have good classification information individually [3]. Feature selection methods try to find the minimal sized subset features that do not decrease the classification accuracy significantly [6]. Several reviews of feature selection techniques are given in early works [3,4,7,8,9]. A typical face recognition system with pixel selection is given in Figure 1.1. This excludes face recognition systems that use partitioning of eye, nose, and lip areas before performing any recognition tasks.

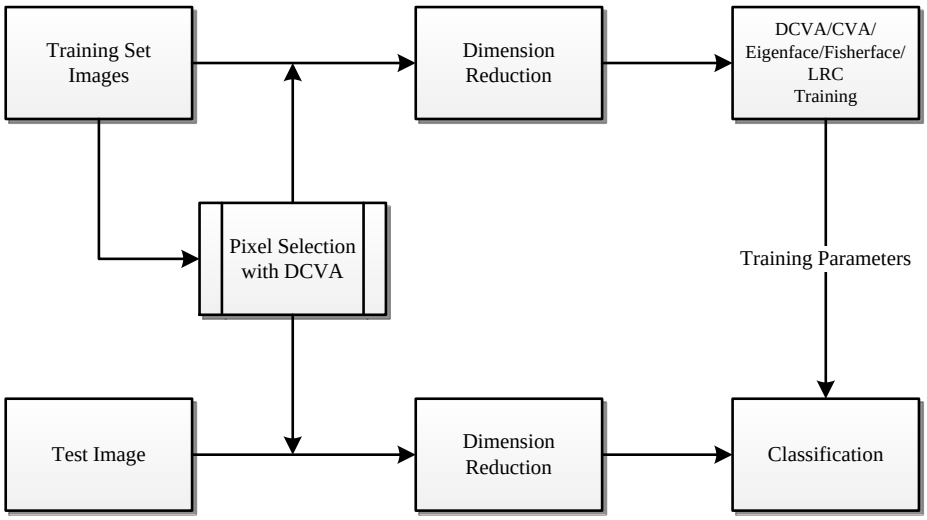


Figure 1.1 The block diagram of a face recognition system [10].

Another factor that affects the performance of the face recognition system is the training sample size [7,11]. One needs sufficient number of training samples to have a well-trained pattern recognition system [12]. The recognition process gets more difficult if only one sample per subject is available which is called *one sample problem* [13]. Traditional methods which use the within-class scatter matrix such as, Fisher linear discriminant analysis (FLDA), Eigenface, Fisherface suffer or fail because the within-class matrix is zero matrix. Several algorithms have been proposed to overcome this challenge [13,14,15,16,17,18]. General tendency at these methods is generating the virtual samples to increase the training set size. But this is not the solution of the singularity problem because in face recognition problems dimension of the feature space is extremely high with respect to the number of feature vectors. This challenge is called *small sample size* problem [4,19]. Various methods are proposed to overcome this difficulty [20,21,22,23,24,25,26]. One method to overcome the singularity problem is using the two dimensional extensions of the one dimensional methods. In [27], two dimensional Fisher linear discriminant analysis (2D-FLDA) with a solution of singularity is proposed. This method is used in [14] and [15] after generating the virtual samples. In [14], image is decomposed using singular value decomposition (SVD). The basis images corresponding to the largest singular values are used to generate the virtual sample. In [15], image and its transpose are decomposed using the QRCP decomposition to generate virtual samples. In [28], we implemented a two dimensional extension of discriminative common vector approach (2D-DCVA) and compare the performances of 1D-DCVA, 2D-FLDA, and 2D-DCVA on different databases. A two dimensional variation of 1D-DCVA is also given in [29] but this method cannot yield unique common vectors which is very important for the performance of the face recognition systems.

In face recognition problems it is also important to determine the distribution of class data. Data of a class can be modeled by its within-class covariance matrix. Sufficient number of data is needed to make good covariance matrix estimation. But especially in face recognition problems this is not possible because of high dimensional and insufficient number of data. In our proposal, we first project the data onto the range space of the total within-class scatter matrix to

reduce the dimensions and try to model a class using not only its own data but also the data of all other classes within this subspace.

In the thesis, we focus on mainly three problems in face recognition; feature selection, one sample problem, and modeling the within-class covariance matrix of the classes in the range subspace of the total within-class scatter matrix. The chapters of the thesis are organized as follows:

In Section 2, the subspace methods used in the thesis are briefly explained. The common vector approach, CVA, was first seen in the speech recognition area [30]. This paper emphasizes the uniqueness of the common vectors. The second important paper is published in the same area in 2001 [31]**.** This paper shows the relation between CVA and the within-class covariance matrix. CVA, alike DCVA, suffers from high dimensional data. In [32], a new implementation of CVA is given to speed up the execution time of the recognition system. A subspace based feature selection method related to CVA was given in [33]. Also many other papers are published that are related to CVA in pattern recognition areas [34,35,36,37,38]. Additionally, we give a novel method to derive CVA using Linear Regression Classification (LRC) [39]. The discriminative common vector approach, DCVA [20], was based on the idea of using the common vector approach for all classes. DCVA has become a popular face recognition method and various papers are published related to DCVA [40,41,42,43,44]. The 2D-FLDA has come into scene due to the need to overcome singularity problem. The 2D-FLDA there is no need to transform the two dimensional images into one dimensional vectors. It estimates the covariance matrices from the image matrices. It overcomes the singularity problem which generally occurs in face recognition problems. Eigenface [45] method is a powerful tool to find a lower dimensional subspace where an image can be represented with negligible loss of information. The image can be perfectly reconstructed using all the eigenfaces that are extracted from the original image. In Fisherface method [21], first PCA [46] is applied to the feature vectors to avoid the singularity of the total within-class scatter matrix. Then LDA [47] is applied to the feature vectors in the reduced space. The LRC based face recognition was first published in 2010 [39]. In this method, it is assumed that the images in a class lie on a linear subspace. Least squares

estimation method is used to determine the regression coefficients. The unknown query is assigned to the class where the minimum reconstruction error occurs.

In Section 3, a novel feature (pixel) selection method is proposed. The features are selected according to the column norms of the projection matrix of the range space of the common vectors. The features corresponding to the columns which have the smallest norms are eliminated first, and then the remaining features are used to form the new feature vector. We also show that importance of the pixels is independent of the selection of the eigenvectors that span the range space projection matrix of the common vectors. We test the performance of the proposed feature selection method in two different face databases with DCVA, CVA, Eigenface, Fisherface, and LRC methods. In the results it is seen that the method successfully eliminates the redundant pixels. We published the method and the results in [10,48]. In [49], a feature selection method based on discriminant features is proposed. Similar to our method, they use the transformation matrices of various feature extraction methods.

In Section 4, a novel image decomposition method which uses the QR decomposition with column pivoting is proposed to overcome the single image problem. Image is decomposed using QRCP method to generate a virtual image. An approximation of the image is generated using the basis images which have the most of the energy of the image. In this section we also gave SVD based image decomposition method [14]. In addition, we propose a two dimensional extension of DCVA. We compare the performances of DCVA, 2D-DCVA, and 2D-FLDA in one sample problem in five different databases. 2D-DCVA clearly outperforms DCVA and 2D-FLDA except in one database. The methods proposed in this section and the experimental results are published in [15,28,50].

In Section 5, two class modeling methods in the range space of the total within-class scatter matrix are proposed. It is observed that in the range space of total within-class scatter matrix, the data of a class is insufficient to model the class which it belongs to. Thus the within-class covariance matrix is singular. In the proposed methods, when we estimate the covariance matrix of a class, we use the data of the both classes. We illustrate the methods with numerical examples

and compare their performances with Support Vector Machines in YALE, ORL, and AR face databases.

The concluding remarks are given in Section 6.

2 SUBSPACE METHODS

Subspace methods are widely used in many pattern recognition applications such as face recognition [2,15,20,21,51,52,53], speech recognition [30,31,37,54,55], handwritten character recognition [56,57,58], spam e-mail detection [59], texture classification [60] etc. In classification problems dimension of the feature vectors are generally high. High dimensional feature vectors may contain redundant or irrelevant features that adversely affect the classification. Subspace methods are successful tools in eliminating the redundancy of the feature vectors and in dimensionality reduction. In subspace methods a transformation is applied to the feature vectors. If suitable transformation is chosen, the transformed feature vector contains most of the discriminatory information in the original features. In this section, brief reviews of Common Vector Approach (CVA), Discriminative Common Vector Approach (DCVA), Two Dimensional Fisher Discriminant Analysis (2D-FLDA), Eigenface, Fisherface, and Linear Regression Classification (LRC) are given.

2.1 Common Vector Approach

A feature vector is considered to have two components: The first one is the component that exhibits properties that are common to the class, and the second one is the remaining component that has all the variations from the common properties. After subtracting the dissimilarities of each vector of a class, the invariant properties of the class will remain. This vector contains the invariant properties, and it is called the common vector. For example, speech signal may contain variances due to personal and environmental effects. In a word class, the differences resulting from the personal and environmental effects are eliminated using the common vector approach(CVA). The residual vector contains the common properties of word class which are unique for each of the word which belongs to that class.

In CVA there are two cases depending on the size of the feature vector(n) and the number of the feature vectors(m): (i) The insufficient case ($n \geq m$), (ii)

The sufficient case ($n < m$). The calculation of common vector is shown in Figure 2.1. CVA divides the space into two orthogonal complementary subspaces, namely *difference* and *indifference*. It is clearly seen from the Figure 2.1 that the common vector lies in the indifference subspace. There are two known methods for calculating the common vector of a class [31]. The first one uses the Gram-Schmidt orthogonalization process, and the second one uses the within-class covariance matrices. In this subsection we give a new implementation of common vectors using LRC.

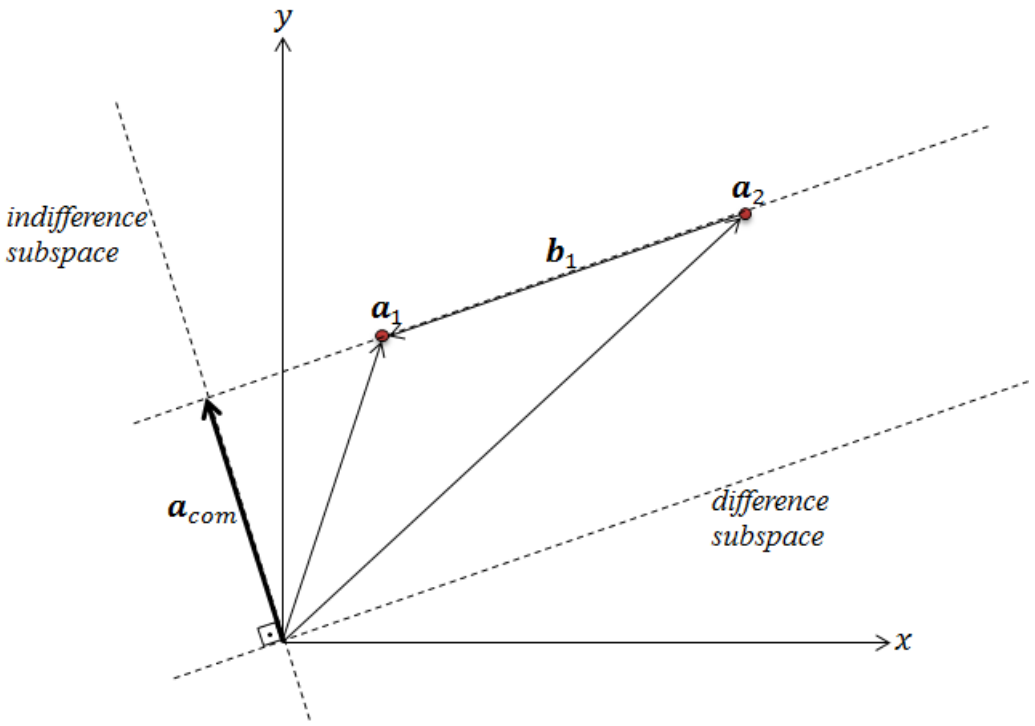


Figure 2.1 Generation of common vector in two dimensional space.

2.1.1 Determining the common vectors from Gram-Schmidt orthogonalization process

Let the training set have C classes and each class has *n*-dimensional *m* samples with $n \geq m$. Let $\boldsymbol{a}_i^j$ be the *n*-dimensional column vector which denotes

the i^{th} sample from the j^{th} class. The difference subspace of j^{th} class is spanned by the difference vectors $\boldsymbol{b}_i^j = \boldsymbol{a}_i^j - \boldsymbol{a}_1^j$, $i = 1, 2, \dots, m$, that is

$$B_j = span\{\boldsymbol{a}_2^j - \boldsymbol{a}_1^j, \dots, \boldsymbol{a}_m^j - \boldsymbol{a}_1^j\} = span\{\boldsymbol{b}_1^j, \dots, \boldsymbol{b}_{m-1}^j\} \tag{2.1}$$

The Gram-Schmidt orthogonalization process [61] is used to obtain the orthonormal vector set $Z = \{\boldsymbol{z}_1^j, \dots, \boldsymbol{z}_{m-1}^j\}$ from the difference vectors $\{\boldsymbol{b}_1^j, \dots, \boldsymbol{b}_{m-1}^j\}$. Additionally Z must satisfy the following equation.

$$\boldsymbol{z}_i^T \boldsymbol{z}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{2.2}$$

The common vector belongs to the j^{th} class can be calculated

$$\boldsymbol{a}_{com}^j = \boldsymbol{a}_i^j - \left(\left(\boldsymbol{a}_i^j\right)^T \cdot \boldsymbol{z}_1^j\right)\boldsymbol{z}_1^j + \cdots + \left(\left(\boldsymbol{a}_i^j\right)^T \cdot \boldsymbol{z}_{m-1}^j\right)\boldsymbol{z}_{m-1}^j \tag{2.3}$$

2.1.2 Determining the common vectors from within-class covariance matrices

Common vector of a class can also be calculated using the within-class covariance matrix. The covariance matrix of the j^{th} class can be calculated as,

$$\boldsymbol{\Phi}_j = \sum_{i=1}^{m} \left(\boldsymbol{a}_i^j - \boldsymbol{a}_{ave}^j\right)\left(\boldsymbol{a}_i^j - \boldsymbol{a}_{ave}^j\right)^T, \tag{2.4}$$

where $\boldsymbol{a}_{ave}^j = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{a}_i^j$.

Covariance matrix characterizes variances of the feature vectors with respect to the average vector. In the insufficient case($n \geq m$), the eigenvectors corresponding to the nonzero eigenvalues form an orthonormal basis to the difference subspace $\boldsymbol{B}$. In this case the indifference subspace $\boldsymbol{B}^{\perp}$, which is the complementary subspace of $\boldsymbol{B}$, is spanned by the eigenvectors corresponding to the zero eigenvalues of covariance matrix. Since the common vector is orthogonal to all vectors in the difference subspace, it must lie in the indifference subspace. Then

the common vector of a class is in the direction of the linear combination of the eigenvectors corresponding to the zero eigenvalues of the covariance matrix[1] or it is the projection of any training data onto the null space of the within –class covariance matrix. The common vector of the j^{th} class can be calculated as

$$\boldsymbol{a}_{com}^{j} = \boldsymbol{P}_{j}^{\perp}\boldsymbol{a}_{i}^{j} = \boldsymbol{a}_{i}^{j} - \boldsymbol{P}_{j}\boldsymbol{a}_{i}^{j}, i = 1, \dots, m, j = 1, \dots, C. \tag{2.5}$$

Here $\boldsymbol{P}_j$ and $\boldsymbol{P}_j^{\perp}$ are the projection matrices of the range space $\boldsymbol{B}$ and the null space $\boldsymbol{B}^{\perp}$ of the covariance matrix and they can be calculated as below

$$\boldsymbol{P}_{j} = \sum_{i=1}^{m-1} \boldsymbol{u}_{i}^{j}\left(\boldsymbol{u}_{i}^{j}\right)^{T} \tag{2.6}$$

$$\boldsymbol{P}_{j}^{\perp} = \sum_{i=m}^{n} \boldsymbol{u}_{i}^{j}\left(\boldsymbol{u}_{i}^{j}\right)^{T}. \tag{2.7}$$

$\{\boldsymbol{u}_1^j, \dots, \boldsymbol{u}_{m-1}^j\}$ are the eigenvectors corresponding to the nonzero eigenvalues and $\{\boldsymbol{u}_m^j, \dots, \boldsymbol{u}_n^j\}$ are the eigenvectors corresponding to the zero eigenvalues.

2.1.3 Determining the common vectors through LRC approach

To apply LRC to CVA, it is better to start with the difference subspace of the j^{th} class. Let $\boldsymbol{a}_1^j, \boldsymbol{a}_2^j, \dots, \boldsymbol{a}_m^j$ be the feature vectors of j^{th} class used in the training stage. The difference subspace of the j^{th} class is spanned by the difference vectors $\{\boldsymbol{a}_2^j - \boldsymbol{a}_1^j, \boldsymbol{a}_3^j - \boldsymbol{a}_1^j \dots, \boldsymbol{a}_m^j - \boldsymbol{a}_1^j\} = \{\boldsymbol{b}_1^j, \boldsymbol{b}_2^j, \dots, \boldsymbol{b}_{m-1}^j\}$. It is known that the subtrahend vector can be any of the feature vector used in the training [30].
The j^{th} distance metric of CVA can be written as

$$d_j = \left|\left(\boldsymbol{a}_x - \boldsymbol{a}_1^j\right) - \widehat{\left(\boldsymbol{a}_x - \boldsymbol{a}_1^j\right)}\right| = \left|\boldsymbol{b}_x - \widehat{\boldsymbol{b}}_x^j\right|. \tag{2.8}$$

$\widehat{\boldsymbol{b}}_x^j$ is the projection of $\boldsymbol{b}_x$ onto the difference subspace of the j^{th} class. If $\boldsymbol{P}_j$ is the projection matrix onto the j^{th} difference subspace, then the above metric will be

$$d_j = \left|\boldsymbol{b}_x - \boldsymbol{P}_j \boldsymbol{b}_x\right|. \tag{2.9}$$

$\boldsymbol{P}_j$ can be determined in similar way that we used in LRC estimation. Let $\boldsymbol{B}_j = \left[\boldsymbol{b}_1^j \vdots \boldsymbol{b}_2^j \vdots \cdots \vdots \boldsymbol{b}_{m-1}^j\right]$ be a matrix whose columns are the difference vectors of the j^{th} class.

$$\widehat{\boldsymbol{b}}_x^j = \boldsymbol{B}_j \boldsymbol{\beta}_{new,j} = \eta_1 \boldsymbol{b}_1^j + \eta_2 \boldsymbol{b}_2^j + \cdots + \eta_{m-1} \boldsymbol{b}_{m-1}^j \tag{2.10}$$

or

$$\widehat{\boldsymbol{b}}_x^j = \boldsymbol{B}_j \boldsymbol{\beta}_{new,j} + \boldsymbol{\varepsilon} \tag{2.11}$$

The sum of error squares is

$$SES = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \left(\boldsymbol{b}_x - \boldsymbol{B}_j \boldsymbol{\beta}_{new,j}\right)^T \left(\boldsymbol{b}_x - \boldsymbol{B}_j \boldsymbol{\beta}_{new,j}\right). \tag{2.12}$$

After the minimization process using least-squares estimation [62,63], vector parameters are obtained as

$$\boldsymbol{\beta}_{new,j} = \left(\boldsymbol{B}_j^T \boldsymbol{B}_j\right)^{-1} \boldsymbol{B}_j^T \boldsymbol{b}_x. \tag{2.13}$$

If we combine (2.8), (2.11) and (2.13), we will get

$$\boldsymbol{P}_j = \boldsymbol{B}_j \left(\boldsymbol{B}_j^T \boldsymbol{B}_j\right)^{-1} \boldsymbol{B}_j^T. \tag{2.14}$$

Let $\boldsymbol{P}_j^{\perp}$ denotes the projection matrix onto the indifference subspace. Then (2.9) becomes

$$\begin{aligned} d_j &= \left|\boldsymbol{b}_x - \boldsymbol{P}_j \boldsymbol{b}_x\right| = \left|\boldsymbol{P}_j^{\perp} \boldsymbol{b}_x\right| \\ &= \left|\boldsymbol{P}_j^{\perp} (\boldsymbol{a}_x - \boldsymbol{a}_1)\right| = \left|\boldsymbol{P}_j^{\perp} \boldsymbol{a}_x - \boldsymbol{P}_j^{\perp} \boldsymbol{a}_1\right| \\ &= \left|\boldsymbol{P}_j^{\perp} \boldsymbol{a}_x - \boldsymbol{a}_{com}^j\right| \end{aligned} \tag{2.15}$$

where $\boldsymbol{a}_{com}^{j}$ is the common vector of the j^{th} class. Here it must be noted that $\boldsymbol{P}_j$ and $\boldsymbol{P}_j^{\perp}$ are idempotent matrices and their sum is equal to identity, that is, $\boldsymbol{P}_j + \boldsymbol{P}_j^{\perp} = \boldsymbol{I}$.

2.1.4 Decision rule

Let $\boldsymbol{a}_x$ be the test sample which will be classified. The projection of $\boldsymbol{a}_x$ onto the indifference subspace of the j^{th} class can be calculated as

$$\boldsymbol{a}_{x,pro} = \boldsymbol{P}_j^{\perp}\boldsymbol{a}_x = \boldsymbol{a}_x - \boldsymbol{P}_j\boldsymbol{a}_x. \tag{2.16}$$

The classification can be done with the following rule.

$$C^* = \underset{j}{\operatorname{argmin}}\{\|\boldsymbol{a}_{x,pro} - \boldsymbol{a}_{com}^{j}\|\}, j = 1, \dots, C. \tag{2.17}$$

2.2 Discriminative Common Vector Approach

Similar to the idea used in the development of CVA, the common vectors in DCVA are obtained using the sum of the within- class scatter matrices. That is, in CVA the common vectors are obtained from each of the within-class scatter matrices whereas in DCVA the common vectors are obtained from the sum of the within class scatter matrices. The common vectors are unique for each of the classes in the training set [20]. An illustration of DCVA for two classes in 3-dimensional case is shown in Figure 2.2. The indifference subspace of these two classes turns out to be the same with the difference subspace of the corresponding common vectors. $\boldsymbol{L}_{com}$ shows the difference subspace of the common vectors. This difference subspace is one dimensional for a two-class case. If all the data that belong to both of the classes are projected onto their indifference subspace first and onto the difference subspace of the common vectors next, then the dimension of all the data will be reduced to one and also 100% recognition rate is guaranteed for the training set.

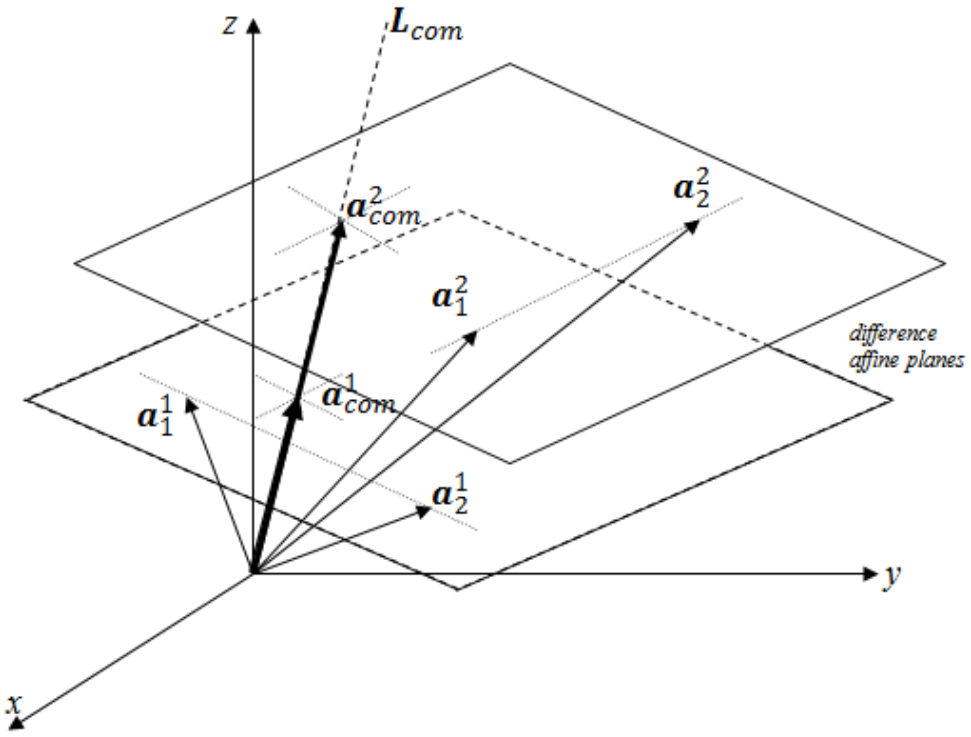


Figure 2.2 An illustration of DCVA in 3-dimensional space for two classes case.

Let C be the number of classes, m be the number of feature vectors in each class, $\boldsymbol{a}_{ij}$ be the i^{th} feature vector of the j^{th} class and let $\boldsymbol{\mu}_j$ be the mean vector of the j^{th} class. The sum of the within-class scatter matrices $\boldsymbol{\Phi}_T$ is defined as

$$\boldsymbol{\Phi}_T = \sum_{j=1}^{C}\sum_{i=1}^{m}(\boldsymbol{a}_i^j - \boldsymbol{\mu}_j)(\boldsymbol{a}_i^j - \boldsymbol{\mu}_j)^T \tag{2.18}$$

The eigenvectors $\boldsymbol{v}_i,\ i = 1,2,\dots,C(m-1)$ belonging to the nonzero eigenvalues of the above total scatter matrix span the difference subspace of all the feature vectors in the training set. Similarly the eigenvectors $\boldsymbol{v}_i,\ i = C(m-1)+1,\dots,n$ belonging to the zero eigenvalues will span the indifference subspace of all the feature vectors in the same training set. Therefore a projection matrix onto this indifference subspace can be formed from the eigenvectors as shown below

$$\boldsymbol{P}^{\perp} = \sum_{i=C(m-1)+1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i^T \tag{2.19}$$

The projections of the feature vectors onto the indifference subspace using this projection matrix yield unique common vectors for each one of the classes [20]. That is,

$$\begin{gathered} \boldsymbol{a}_{com}^{j} = \boldsymbol{P}^{\perp}\boldsymbol{a}_i^{j}, \;\; i = 1, \dots, m, \;\; j = 1, \dots, C \\ \boldsymbol{a}_{com}^{j} \neq \boldsymbol{a}_{com}^{k} \text{ if } j \neq k \end{gathered} \tag{2.20}$$

After obtaining the common vectors of all the classes in the database, a difference subspace of the common vectors can be obtained. As before the orthonormal basis vectors $\boldsymbol{e}_i, \; i = 1, \dots, C-1$ of this subspace can be obtained either from the Gram-Schmidt orthogonalization process or from the scatter matrix of the common vectors.

If one wants to use the scatter matrix approach, then the scatter matrix of common vectors must be calculated first,

$$\boldsymbol{\Phi}_{com} = \sum_{i=1}^{C} \left(\boldsymbol{a}_{com}^{i} - \boldsymbol{\mu}_{com}\right)\left(\boldsymbol{a}_{com}^{i} - \boldsymbol{\mu}_{com}\right)^T \tag{2.21}$$

where $\boldsymbol{\mu}_{com} = \frac{1}{C}\sum_{i=1}^{C} \boldsymbol{a}_{com}^{i}$ is the mean of the common vectors. Then the projection matrix $\boldsymbol{W}$ onto the difference subspace of the common vectors must be obtained, that is, $\boldsymbol{W} = [\boldsymbol{w}_1 \vdots \boldsymbol{w}_2 \vdots \cdots \vdots \boldsymbol{w}_{C-1}]^T$ can be obtained using the eigenvectors $\boldsymbol{w}_i, \; i = 1,2, \dots, C-1$ corresponding to the nonzero eigenvalues of $\boldsymbol{\Phi}_{com}$.

Since the distances between the common vectors in the whole space are kept the same as the distances they have in the difference subspace of the common vectors, the classification rule can be written in the difference subspace of the common vectors. This will reduce the dimensions used in the classification phase

from n to $C-1$. Remembering that n is the whole space dimension and C is the number of classes in the database; this is an important reduction in dimensionality.

2.2.1 Decision rule

In the classification phase let $\boldsymbol{a}_x$ be the test vector which will be classified, then the classification can be done by the following

$$C^* = \underset{j}{\operatorname{argmin}}\{\|\boldsymbol{W}(\boldsymbol{a}_i^j - \boldsymbol{a}_x)\|\}, \qquad j = 1, \dots, C \tag{2.22}$$

where $\boldsymbol{W}$ is a $(C-1) \times n$ dimensional projection matrix onto the difference subspace of the common vectors.

2.3 Two Dimensional Fisher Linear Discriminant Analysis

The objective of FLDA is to perform dimensionality reduction while preserving as much information as possible. LDA allows one to choose a linear subspace of the original feature space in a way that maximizes the between-class variance with respect to the within-class variance [64]. But in face recognition problems generally the dimension of the feature vectors are very high with respect to the number of feature vectors. Since within-class scatter matrix has zero eigenvalues, the FLDA cannot be directly applied. To overcome this challenge, 2D feature matrices are used instead of one dimensional feature vectors in [27,65]. Also 2D-FLDA generally outperforms 1D-FLDA [66,67]. The 2D-FLDA can be summarized as follows:

Let C be the number of classes, N be the number of selected samples from each class, A_j^i be the j^{th} image from i^{th} class and M^i be the average image of i^{th} class, that is,

$$\boldsymbol{M}^i = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{A}_j^i, j = 1, \dots, C \tag{2.23}$$

In 2D-FLDA, the optimal projection vectors $\boldsymbol{X} = [\boldsymbol{x}_1 \vdots \boldsymbol{x}_2 \vdots \cdots \vdots \boldsymbol{x}_d]$ are to be found. Here d is *at most* $\min(C-1, n)$. The optimal projection vectors can be calculated by maximizing the following criterion

$$J(X) = \frac{\boldsymbol{X}^T S_B \boldsymbol{X}}{\boldsymbol{X}^T S_W \boldsymbol{X}} \tag{2.24}$$

where

$$\boldsymbol{S}_W = \sum_{i=1}^{C} \sum_{j=1}^{N} \left(\boldsymbol{A}_j^i - \boldsymbol{M}^i\right)^T \left(\boldsymbol{A}_j^i - \boldsymbol{M}^i\right) \tag{2.25}$$

$$\boldsymbol{S}_B = \sum_{i=1}^{C} \left(\boldsymbol{M}^i - \boldsymbol{M}\right)^T \left(\boldsymbol{M}^i - \boldsymbol{M}\right) \tag{2.26}$$

$$\boldsymbol{M} = \frac{1}{C} \sum_{i=1}^{C} \boldsymbol{M}^i \tag{2.27}$$

2.3.1 Decision rule

Let $\boldsymbol{A}_j^i, i = 1, \dots, C, j = 1, \dots, N$ be the j^{th} image from i^{th} class and let $\boldsymbol{A}_{test}$ be the unknown image that will be classified. Then the optimal projection vectors for $\boldsymbol{A}_j^i$ and $\boldsymbol{A}_{test}$ can be calculated as;

$$\boldsymbol{B}_j^i = \boldsymbol{A}_j^i[\boldsymbol{x}_1 \vdots \boldsymbol{x}_2 \vdots \cdots \vdots \boldsymbol{x}_d] = \left[\boldsymbol{y}_1^{i,j} \vdots \boldsymbol{y}_2^{i,j} \vdots \cdots \vdots \boldsymbol{y}_d^{i,j}\right], i = 1, \dots, C, \quad j = 1, \dots, N \tag{2.28}$$

and

$$\boldsymbol{B}_{test} = \boldsymbol{A}_{test}[\boldsymbol{x}_1 \vdots \boldsymbol{x}_2 \vdots \cdots \vdots \boldsymbol{x}_d] = [\boldsymbol{y}_1^{test} \vdots \boldsymbol{y}_2^{test} \vdots \cdots \vdots \boldsymbol{y}_d^{test}] \tag{2.29}$$

respectively. $\boldsymbol{A}_{test}$ is assigned to the class i where the nearest $\boldsymbol{B}_j^i$ belongs to, using the following metric:

$$C^* = \underset{i}{\operatorname{argmin}}\left\{D\left(\boldsymbol{A}_{test}, \boldsymbol{A}_j^i\right)\right\}, i = 1, \dots, C, j = 1, \dots, N \tag{2.30}$$

$$= \underset{i}{\operatorname{argmin}}\left\{D\left(\boldsymbol{B}_{test}, \boldsymbol{B}_j^i\right)\right\}, i = 1, \dots, C, j = 1, \dots, N \tag{2.31}$$

$$= \underset{i}{\operatorname{argmin}}\left\{\sum_{k=1}^{d}\left\|\boldsymbol{y}_k^{test} - \boldsymbol{y}_k^{i,j}\right\|_2\right\}, i = 1, \dots, C, j = 1, \dots, N \tag{2.32}$$

where $\|\cdot\|_2$ denotes the Euclidean distance between two vectors.

2.4 Eigenface

Eigenface method uses PCA method to reduce the dimension. It finds the eigenvectors $W = [\boldsymbol{w}_1 \vdots \boldsymbol{w}_2 \vdots \cdots \vdots \boldsymbol{w}_d]$ that maximizes the objective function

$$J(\boldsymbol{W}) = |\boldsymbol{W}^T \boldsymbol{\Phi}_T \boldsymbol{W}| \tag{2.33}$$

where $\boldsymbol{\Phi}_{\mathrm{T}}$ is the total scatter matrix defined by

$$\boldsymbol{\Phi}_{\mathrm{T}} = \sum_{i=1}^{N}\left(\boldsymbol{a}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{a}_i - \boldsymbol{\mu}\right)^T \tag{2.34}$$

where $\boldsymbol{\mu}$ is the mean of the all samples. $\boldsymbol{w}_i, i = 1,2, \dots, d$ are the eigenvectors corresponding to the largest eigenvalues of $\boldsymbol{\Phi}_T$.

The classification is done using nearest neighbor classifier in the transformed space. Let $\boldsymbol{a}_{test}$ be the unknown image that will be classified. $\boldsymbol{a}_{test}$ can be projected into the eigenspace by using the following equation

$$\boldsymbol{\Omega}_{test} = \boldsymbol{W}^T(\boldsymbol{a}_{test} - \boldsymbol{\mu}) \tag{2.35}$$

Similarly $\boldsymbol{a}_i$ is represented in the eigenspace by the $\boldsymbol{\Omega}_i = \boldsymbol{W}^T(\boldsymbol{a}_i - \boldsymbol{\mu})$. Then $\boldsymbol{a}_{test}$ is classified to the class according to the following criterion,

$$\varepsilon = \underset{i}{\text{argmin}}\{\|\boldsymbol{\Omega}_{test} - \boldsymbol{\Omega}_i\|\}. \tag{2.36}$$

In addition to the above criterion, if ε is greater than predetermined threshold θ_ε , then $\boldsymbol{a}_{test}$ is assigned to an unknown face.

2.5 Fisherface

It is known that in face recognition problems $\boldsymbol{S_W}$ is singular since the number of feature vectors is much smaller than the number of feature vectors . In Fisherface method [21,23] the feature vector space is reduced by using PCA to avoid the singularity problem of $\boldsymbol{S_W}$. Subsequently, LDA is applied to the feature vectors in the reduced space. Let a_j^i be the j^{th} feature vector from i^{th} class, $\boldsymbol{\mu}_i$ be the mean vector of the i^{th} class. The within-class covariance matrix and the between-class covariance matrix are defined as

$$\boldsymbol{S}_W = \sum_{i=1}^{C}\sum_{j=1}^{N}(\boldsymbol{a}_j^i - \boldsymbol{\mu}^i)(\boldsymbol{a}_j^i - \boldsymbol{\mu}^i)^T \tag{2.37}$$

$$\boldsymbol{S}_B = \sum_{i=1}^{C}(\boldsymbol{\mu_i} - \boldsymbol{\mu})(\boldsymbol{\mu_i} - \boldsymbol{\mu})^T \tag{2.38}$$

where and $\boldsymbol{\mu} = 1/C\sum_{i=1}^{C}\boldsymbol{\mu}_i$ is the mean of mean vectors of the classes. The optimization can be done by maximizing the criterion given in the below:

$$J(\boldsymbol{W}) = \frac{\left|\boldsymbol{W}^T\boldsymbol{W}_{\text{pca}}^T\mathbf{S}_B\boldsymbol{W}_{\text{pca}}\boldsymbol{W}\right|}{\left|\boldsymbol{W}^T\boldsymbol{W}_{\text{pca}}^T\mathbf{S}_W\boldsymbol{W}_{\text{pca}}\boldsymbol{W}\right|} \tag{2.39}$$

Here $\boldsymbol{W}_{\mathbf{pca}}$ is chosen to maximize the criterion $\boldsymbol{J}(\boldsymbol{W}_{\text{pca}}) = |\boldsymbol{W}^T\boldsymbol{\Phi}_\mathbf{T}\boldsymbol{W}|$ where $\boldsymbol{\Phi}_\mathbf{T}$ is given in (2.34).

2.6 Linear Regression Classification

Let C be the number of classes, m be the number of feature vectors of a class used in training, and $\{\boldsymbol{a}_1^j, \boldsymbol{a}_2^j, \dots, \boldsymbol{a}_m^j\}$ be the feature vectors of the training set of the j^{th} class. LRC idea is based on a distance metric given by

$$d_j = \left|\boldsymbol{a}_x - \boldsymbol{a}_x^j\right| \tag{2.40}$$

where $\boldsymbol{a}_x$ is the test feature vector and $\boldsymbol{a}_x^j$ is its projection onto the subspace spanned by the feature vectors of the training set of the j^{th} class. Let $\boldsymbol{P}_j$ be the projection matrix onto the j^{th} class, then the distance metric will become

$$d_j = \left|\boldsymbol{a}_x - \boldsymbol{P}_j\boldsymbol{a}_x\right|. \tag{2.41}$$

$\boldsymbol{P}_j$ is calculated using a linear combination of the feature vectors of the training set of the j^{th} class under a constraint relation in its optimized form in terms of minimum sum of error squares

$$\boldsymbol{a}_x^j = \boldsymbol{W}_j\boldsymbol{\beta}_j \tag{2.42}$$

with respect to the coefficients $\boldsymbol{\beta}_j$. $\boldsymbol{W}_j$ is a matrix formed from the feature vectors in the training set of the j^{th} class, i.e., $\boldsymbol{W}_j = \left[\boldsymbol{a}_1^j \vdots \boldsymbol{a}_2^j \vdots \cdots \vdots \boldsymbol{a}_m^j\right]$.
The projection of a feature vector onto the i^{th} class can be calculated from

$$\boldsymbol{a}_x = \boldsymbol{W}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon} \tag{2.43}$$

where $\boldsymbol{\varepsilon}$ is the error or the remaining part of $\boldsymbol{a}_x$ in the rest of the whole space $\mathbb{R}^n$ of feature vectors.
The sum of the error squares can be formed with ease

$$SES = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \left(\boldsymbol{a}_x - \boldsymbol{W}_j\boldsymbol{\beta}_j\right)^T\left(\boldsymbol{a}_x - \boldsymbol{W}_j\boldsymbol{\beta}_j\right). \tag{2.44}$$

To minimize *SES*, the critical point(s) must be calculated by taking the derivative of *SES* with respect to $\boldsymbol{\beta}_j$.

$$\frac{\partial SES}{\partial \boldsymbol{\beta}_j} = \mathbf{0} \tag{2.45}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}_j}\left[\boldsymbol{a}_x^T\boldsymbol{a}_x - \boldsymbol{\beta}_j^T\boldsymbol{W}_j^T\boldsymbol{a}_x - \boldsymbol{a}_x^T\boldsymbol{W}_j\boldsymbol{\beta}_j + \boldsymbol{\beta}_j^T\boldsymbol{W}_j^T\boldsymbol{W}_j\boldsymbol{\beta}_j\right] = \mathbf{0} \tag{2.46}$$

$$-2\boldsymbol{W}_j^T\boldsymbol{a}_x + 2\boldsymbol{W}_j^T\boldsymbol{W}_j\boldsymbol{\beta}_j = \mathbf{0} \tag{2.47}$$

$$\boldsymbol{\beta}_j = \left(\boldsymbol{W}_j^T\boldsymbol{W}_j\right)^{-1}\boldsymbol{W}_j^T\boldsymbol{a}_x \tag{2.48}$$

Then the distance metric of LRC becomes

$$d_j = \left|\boldsymbol{a}_x - \boldsymbol{a}_x^j\right| = \left|\boldsymbol{a}_x - \boldsymbol{W}_j\boldsymbol{\beta}_j\right| = \left|\boldsymbol{a}_x - \boldsymbol{W}_j\left(\boldsymbol{W}_j^T\boldsymbol{W}_j\right)^{-1}\boldsymbol{W}_j^T\boldsymbol{a}_x\right|. \tag{2.49}$$

Also from (2.41) the projection matrix to the j^{th} class $\boldsymbol{P}_j = \boldsymbol{W}_j\left(\boldsymbol{W}_j^T\boldsymbol{W}_j\right)^{-1}\boldsymbol{W}_j^T$ is obtained. If the orthogonal complement of the projection matrix $\boldsymbol{P}_j$ is shown with $\boldsymbol{P}_j^{\perp}$, then the metric given in (2.49) is equal to

$$d_j = \left|\boldsymbol{P}_j^{\perp}\boldsymbol{a}_x\right| \tag{2.50}$$

Here $\boldsymbol{P}_j^{\perp}\boldsymbol{a}_x$ is the projection of $\boldsymbol{a}_x$ onto the subspace that complements the subspace of the j^{th} class with $\boldsymbol{P}_j + \boldsymbol{P}_j^{\perp} = \boldsymbol{I}$ to the whole space. Therefore d_j is the distance of $\boldsymbol{a}_x$ to the subspace formed from the feature vectors of the j^{th} class. In that sense if $\boldsymbol{a}_x$ belongs to the j^{th} class, $\boldsymbol{P}_j^{\perp}\boldsymbol{a}_x$ must be negligibly small, or else $\boldsymbol{P}_j\boldsymbol{a}_x$ is almost equal to $\boldsymbol{a}_x$. Let

$$\boldsymbol{a}_x^j = \boldsymbol{W}_j\left(\boldsymbol{W}_j^T\boldsymbol{W}_j\right)^{-1}\boldsymbol{W}_j^T\boldsymbol{a}_x = \boldsymbol{P}_j\boldsymbol{a}_x, \qquad j = 1,2,\dots,C \tag{2.51}$$

be the predicted response vectors. Classification is done according to the following decision rule.

$$C^* = \underset{j}{\operatorname{argmin}} \left\{ \left\| \boldsymbol{a}_x - \boldsymbol{a}_x^j \right\|_2 \right\}, j = 1,2, \dots, C \tag{2.52}$$

3 A NOVEL FEATURE SELECTION METHOD IN THE RANGE SPACE OF COMMON VECTORS

In face recognition problems dimensions of the feature vectors are generally high. This increases computational complexity and execution time of the system and therefore degrades the performance of the system. However there might be redundant features which increase the computational cost while increasing the recognition accuracy negligibly.

In this chapter we introduce a new feature selection algorithm using the projection matrix of the common vectors. We presented the primary outcomes of our work in [48], secondly we performed more detailed work in [10]. In this work, a novel feature selection algorithm is introduced related to DCVA. The importance of the pixels is determined by the column norms of the projection matrix of the common vectors of all the classes. Then the pixels corresponding to the columns which have the smallest norms are omitted since their contribution to the classification criteria will be negligible. The pixels corresponding to the columns which have the largest norms are used in constructing the reduced dimensional feature vectors. Since the dimension of the reduced feature vector is less than the original one, storage and speed improvements are achieved which are important for real-time and real-life applications. It is also shown that the order of the magnitudes of the column vector norms is not affected from the choices of the orthonormal basis vector set that span the range space of the common vectors.

3.1 Determining the Importance of Features Using the Projection Matrix of the Range Space of the Common Vectors

It is assumed that not all the gray level values from all pixels have equal importance in the classification of the face images. Depending on their importance, some of the features or equivalently pixels can be eliminated from the face images without having much effect on the classification rates. For the realization of this idea, the transformation or the projection matrix $\boldsymbol{W}$ is used since it is a projection matrix that transforms the original n – dimensional training data

onto a $C-1$ dimensional difference subspace of the common vectors. This means that the projection matrix $\boldsymbol{W}$ is obtained from the within–class scatter matrix of the common vectors $\boldsymbol{\Phi}_{com}$. Because of this $C-1$ dimensional difference subspace, the transformation matrix $\boldsymbol{W}$ must have major importance in dimensionality reduction. Each one of the columns of $\boldsymbol{W}$ is multiplied with the gray level value of the corresponding pixel of the face image as it is given in (3.3). This will be explained with the row and the column vectors of the matrix $\boldsymbol{W}$.

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \vdots \\ \boldsymbol{w}_{C-1}^T \end{bmatrix} \tag{3.1}$$

where $\boldsymbol{w}_i$ for $i = 1, \dots, C-1$ are basis vectors that span the difference subspace of $\boldsymbol{\Phi}_{com}$. More explicitly $\boldsymbol{W}$ is a $(C-1) \times n$ dimensional matrix whose elements are given below

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & & w_{2n} \\ \vdots & & \ddots & \vdots \\ w_{(C-1)1} & w_{(C-1)2} & \cdots & w_{(C-1)n} \end{bmatrix} \tag{3.2}$$

In (3.2), if i^{th} column of $\boldsymbol{W}$ is called $\boldsymbol{q}_i$, which is a $(C-1) \times 1$ vector, then $\boldsymbol{W}$ can be expressed with its column vectors as $\boldsymbol{W} = [\boldsymbol{q}_1 \vdots \boldsymbol{q}_2 \vdots \cdots \vdots \boldsymbol{q}_n]$. Let $\boldsymbol{a} = [a_1 \quad a_2 \quad \cdots \quad a_n]^T$ be an n – dimensional face image vector where each $a_i,\ i = 1, \dots, n$ corresponds to a gray level of a pixel in the face image. From here on, a_i will be called the i^{th} feature of the face image vector. Then

$$\boldsymbol{W}\boldsymbol{a} = [\boldsymbol{q}_1 \vdots \boldsymbol{q}_2 \vdots \cdots \vdots \boldsymbol{q}_n]\boldsymbol{a} = \boldsymbol{q}_1 a_1 + \boldsymbol{q}_2 a_2 + \cdots \boldsymbol{q}_n a_n \tag{3.3}$$

$$= \begin{bmatrix} w_{11} \\ w_{21} \\ \vdots \\ w_{(C-1)1} \end{bmatrix} a_1 + \begin{bmatrix} w_{12} \\ w_{22} \\ \vdots \\ w_{(C-1)2} \end{bmatrix} a_2 + \cdots + \begin{bmatrix} w_{1n} \\ w_{2n} \\ \vdots \\ w_{(C-1)n} \end{bmatrix} a_n \tag{3.4}$$

In the above equation it is seen that the projection of any face image vector onto the difference subspace of the common vectors is the sum of the multiplication of the column vectors of $\boldsymbol{W}$ with the gray levels of each one of the

pixels. That is, the elements of $\boldsymbol{Wa}$ are linear combinations of the features of the face image vector, $\boldsymbol{a}$. This may give some hint about which pixels contribute more. After all, $\boldsymbol{Wa}$ is directly used in the classification criteria as it is given in (2.22).

In equation (3.4), it can be easily seen that the importance of i^{th} feature a_i of the face image vector $\boldsymbol{a}$ is strongly related with the norm of the i^{th} column vector $\|\boldsymbol{z}_i\|$. If all elements in the i^{th} column of $\boldsymbol{W}$ are zero, then the i^{th} element of the feature vector $\boldsymbol{a}$ has no effect on the projected face image vector. Then the i^{th} element or feature can be eliminated. If the number of classes is small, it may be possible to find all zeros in some columns of the transformation matrix. But in high dimensional spaces it is expected that it will be hard to find such zero columns. Even in this case, the magnitudes of the norms of the columns of $\boldsymbol{W}$ can be used in feature selection. The column that has a small norm will have negligible effect on the calculations of the projected face image vectors. The following numerical example illustrates the proposed pixel elimination method.

Example 3.1 Let $C_1 = \left\{ \begin{bmatrix} 3 \\ 2 \\ 1 \\ 2 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 \\ 3 \\ -2 \\ 2 \\ 2 \\ 3 \end{bmatrix} \right\}$, $C_2 = \left\{ \begin{bmatrix} 2 \\ 4 \\ 4 \\ 1 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 \\ 8 \\ -4 \\ 0 \\ 3 \\ -1 \end{bmatrix} \right\}$, and $C_3 = \left\{ \begin{bmatrix} 4 \\ 1 \\ -4 \\ -1 \\ -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ -3 \\ 1 \\ 4 \\ -2 \end{bmatrix} \right\}$ are three classes and $\boldsymbol{a}_{test} = [2 \quad 5 \quad 4 \quad 3 \quad 1]^T$ is the unknown feature vector which is to be classified.

The projection matrix given in Eq.(12) is calculated as

$$\boldsymbol{W} = \begin{bmatrix} -0.5574 & 0.1751 & -0.0592 & -0.7368 & 0.0496 & 0.3314 \\ -0.1678 & -0.8642 & -0.4641 & -0.0693 & -0.0397 & -0.0565 \end{bmatrix} \tag{3.5}$$

The projection of the test vector onto the range space of $\boldsymbol{W}$ can be calculated as

$$\boldsymbol{W}\boldsymbol{a}_{test} = \begin{bmatrix} -0.5574 & 0.1751 & -0.0592 & -0.7368 & 0.0496 & 0.3314 \\ -0.1678 & -0.8642 & -0.4641 & -0.0693 & -0.0397 & -0.0565 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 4 \\ 2 \\ 3 \\ 1 \end{bmatrix} \quad (3.6)$$

$$= \begin{bmatrix} -0.5574 \\ -0.1678 \end{bmatrix} 2 + \begin{bmatrix} 0.1751 \\ -0.8642 \end{bmatrix} 5 + \begin{bmatrix} -0.0592 \\ -0.4641 \end{bmatrix} 4 + \begin{bmatrix} -0.7368 \\ -0.0693 \end{bmatrix} 2 + \begin{bmatrix} 0.0496 \\ -0.0397 \end{bmatrix} 3 + \begin{bmatrix} 0.3314 \\ -0.0565 \end{bmatrix} 1 \quad (3.7)$$

$$= \begin{bmatrix} -1.1148 \\ -0.3356 \end{bmatrix} + \begin{bmatrix} 0.8755 \\ -4.3210 \end{bmatrix} + \begin{bmatrix} -0.2368 \\ -1.8564 \end{bmatrix} + \begin{bmatrix} -1.4736 \\ -0.1368 \end{bmatrix} + \begin{bmatrix} 0.1488 \\ -0.1191 \end{bmatrix} + \begin{bmatrix} 0.3314 \\ -0.0565 \end{bmatrix} \quad (3.8)$$

$$= \begin{bmatrix} -1.4695 \\ -6.8254 \end{bmatrix} \quad (3.9)$$

The norms of the columns of $\boldsymbol{W}$ are 0.5821, 0.8818, 0.4678, 0.7400, 0.0635 and 0.3362 respectively. The 5th column of $\boldsymbol{W}$ has the smallest norm. If we calculate $\boldsymbol{W}\boldsymbol{a}_{test}$ by ignoring the 5th element of the feature vectors, the projection matrix in the reduced space becomes

$$\boldsymbol{W}_{red} = \begin{bmatrix} -0.6236 & 0.2084 & -0.1055 & -0.6528 & 0.3611 \\ -0.1195 & -0.8828 & -0.4249 & -0.1413 & -0.0765 \end{bmatrix}. \quad (3.10)$$

The projection of the reduced test vector onto the range space of $\boldsymbol{W}_{red}$ is $\boldsymbol{W}_{red}\boldsymbol{a}_{test}^{red} = \begin{bmatrix} -1.5721 \\ -6.7117 \end{bmatrix}$. The distances of the test vector to the discriminative common vectors of the classes in (2.22) are 4.0749, 1.2472, and 7.2667 respectively, then $\boldsymbol{a}_{test}$ is classified to C_2 according to the criteria given in Eq(10). Similarly, the distances of $\boldsymbol{a}_{test}^{red}$ to the discriminative common vectors of the classes in (2.22) are 3.9163, 1.2437, and 7.1520 respectively, then $\boldsymbol{a}_{test}^{red}$ is classified to C_2. The $\boldsymbol{a}_{test}^{red}$ is classified to the class C_2 as it is with the previous case of $\boldsymbol{a}_{test}$. The 5th dimension or the fifth element of the feature vector can be thought as a redundant feature.

Since the basis vector set that spans a subspace will not be ever unique, the next question would be that do the importance levels of the features (or

equivalently pixels) change depending on the choice of the basis vector set? Considering this situation, it will be shown that even though the basis vector set that forms the transformation matrix $\boldsymbol{W}$ is not unique, the order of the magnitudes of the column vector norms does not change. Depending on this fact, the importance level of the pixels will not change with different choices of the basis vector set. This is given by the following theorem.

Theorem: Let $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_{C-1}\}$ and $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_{C-1}\}$ be two different orthonormal basis vector sets that span the range space of S_{com}. Obtaining the transformation matrix $\boldsymbol{W}$ using both of these orthonormal basis vector sets does not change the order of the magnitudes of the column vector norms of $\boldsymbol{W}$.

Proof: Let $\boldsymbol{W} = [\boldsymbol{w}_1 \vdots \boldsymbol{w}_2 \vdots \cdots \vdots \boldsymbol{w}_{C-1}]^T$ and $\boldsymbol{V} = [\boldsymbol{v}_1 \vdots \boldsymbol{v}_2 \vdots \cdots \vdots \boldsymbol{v}_{C-1}]^T$ be the transformation matrices of range space of $\boldsymbol{S}_{com}$, where $\boldsymbol{w}_i^T$, $\boldsymbol{v}_i^T$ for $i = 1,2, \dots, C-1$ are the row vectors of $\boldsymbol{W}$ and $\boldsymbol{V}$ matrices, respectively.
We can rewrite the transformation matrices with the column vectors as before

$$\boldsymbol{W} = [\boldsymbol{q}_1 \vdots \boldsymbol{q}_2 \vdots \cdots \vdots \boldsymbol{q}_n], \qquad \boldsymbol{V} = [\boldsymbol{p}_1 \vdots \boldsymbol{p}_2 \vdots \cdots \vdots \boldsymbol{p}_n] \tag{3.11}$$

where $\boldsymbol{q}_i$ and $\boldsymbol{p}_i, i = 1, \dots, n$ are $C-1$ dimensional column vectors of the $\boldsymbol{W}$ and $\boldsymbol{V}$ matrices respectively. Since the columns of $\boldsymbol{W}$ and $\boldsymbol{V}$ span the range space of $\boldsymbol{S}_{com}$, the following equation holds [20].

$$\boldsymbol{P} = \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{V}^T\boldsymbol{V} \tag{3.12}$$

where $\boldsymbol{P}$ is the projection matrix onto the difference subspace of the common vectors.

$$\begin{bmatrix} \boldsymbol{q}_1^T\boldsymbol{q}_1 & \boldsymbol{q}_1^T\boldsymbol{q}_2 & \cdots & \boldsymbol{q}_1^T\boldsymbol{q}_n \\ \boldsymbol{q}_2^T\boldsymbol{q}_1 & \boldsymbol{q}_2^T\boldsymbol{q}_2 & & \boldsymbol{q}_2^T\boldsymbol{q}_n \\ \vdots & & \ddots & \vdots \\ \boldsymbol{q}_n^T\boldsymbol{q}_1 & \boldsymbol{q}_n^T\boldsymbol{q}_2 & \cdots & \boldsymbol{q}_n^T\boldsymbol{q}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}_1^T p_1 & \boldsymbol{p}_1^T\boldsymbol{p}_2 & \cdots & \boldsymbol{p}_1^T\boldsymbol{p}_n \\ \boldsymbol{p}_2^T\boldsymbol{p}_1 & \boldsymbol{p}_2^T\boldsymbol{p}_2 & & \boldsymbol{p}_2^T\boldsymbol{p}_n \\ \vdots & & \ddots & \vdots \\ \boldsymbol{p}_n^T\boldsymbol{p}_1 & \boldsymbol{p}_n^T\boldsymbol{p}_2 & \cdots & \boldsymbol{p}_n^T\boldsymbol{p}_n \end{bmatrix} \tag{3.13}$$

Then the corresponding elements in the two matrices are equal, that is,

$$\boldsymbol{q}_i^T \boldsymbol{q}_i = \boldsymbol{p}_i^T \boldsymbol{p}_i, \;\; i = 1, \dots, n \tag{3.14}$$

$$\|\boldsymbol{q}_i\|^2 = \|\boldsymbol{p}_i\|^2, \;\; i = 1, \dots, n \tag{3.15}$$

$$\|\boldsymbol{q}_i\| = \|\boldsymbol{z}_i\|, \;\; i = 1, \dots, n \tag{3.16}$$

This means that the norms of the column vectors of $\boldsymbol{W}$ will not change with different choices of the basis vector set. This completes the proof.

In Figure 3.1, the column norms of $\boldsymbol{W}$ in the sorted order are shown. Eliminating the columns of $\boldsymbol{W}$ which have the smallest norms eliminates corresponding pixels in the image. Figure 3.2 shows the original image and the pixel reduced images using the proposed feature selection method. Here eliminated pixels are shown in green color. In this figure, it can be easily seen that eye, nose, and lips remain, which means that they have the most important discriminatory information.

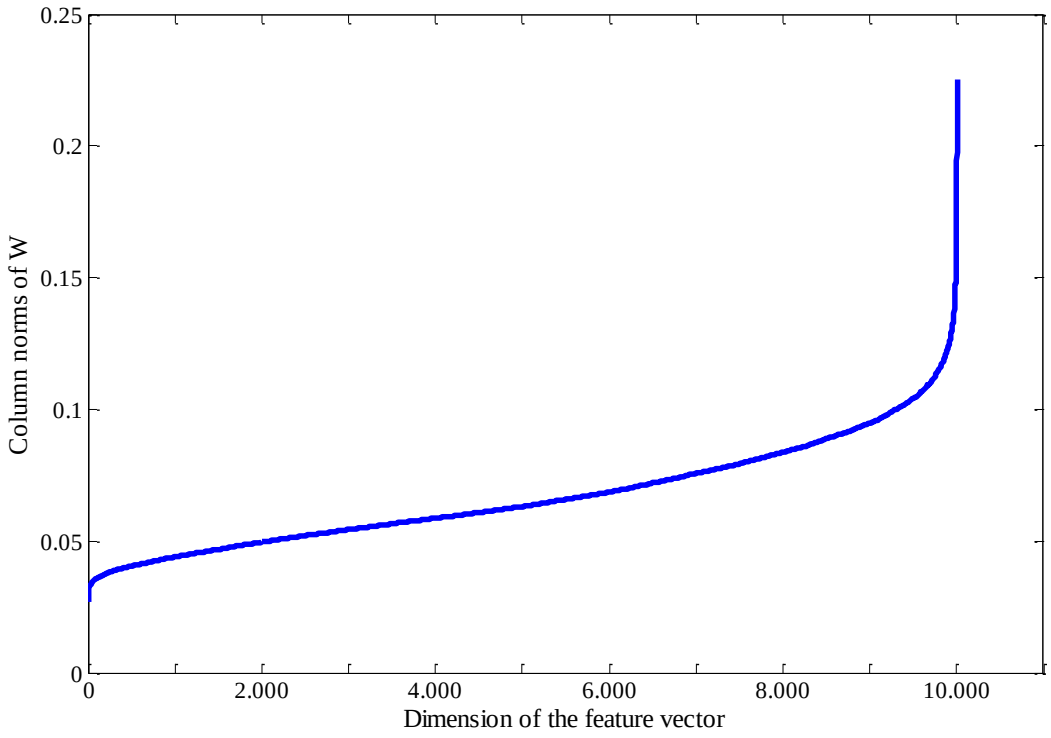
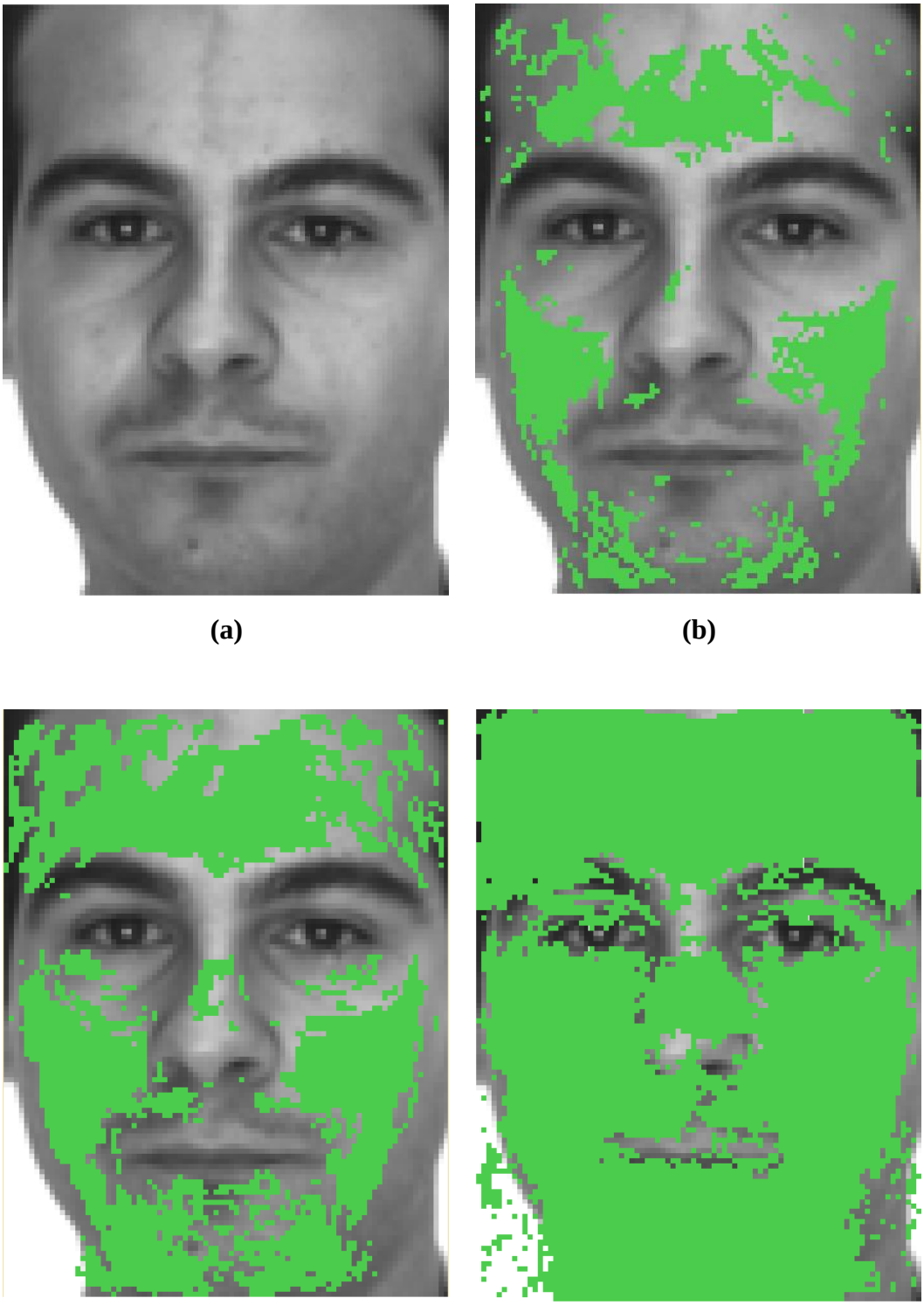


Figure 3.1 The column norms of W in the ascending order



(a) **(b)**

(c) **(d)**

Figure 3.2 (a) Original image and (b) 2000, (c) 4000, and (d) 8000 pixels eliminated images, using the proposed feature selection method.

3.2 Experimental Work

In the experimental work, the pixels are selected using the training set images in accordance with the method mentioned in Section 3.1. Then the images with the reduced number of pixels of the training set are used for training purposes. In the experimental stage we used two different databases, AR [68], ORL [69], and YALE [21]. We compare the recognition performances of DCVA, CVA, Eigenface, Fisherface, and LRC methods.

3.2.1 AR face database

The AR database contains 126 subjects with 26 images with different frontal views taken under different illumination conditions, occlusions and facial expressions. The original size of images is 768×576 pixels. The same subjects are used from AR database mentioned in reference [20]. This database contains 50 people of which 30 are male and 20 are female. After the aligning, scaling, localizing, cropping, and resizing operations, the final size of the images was 115×87 pixels so that 10,005 dimensional feature vectors are obtained from the face images. In Figure 3.3 the images of a subject from the database after the mentioned preprocessing operations are shown.

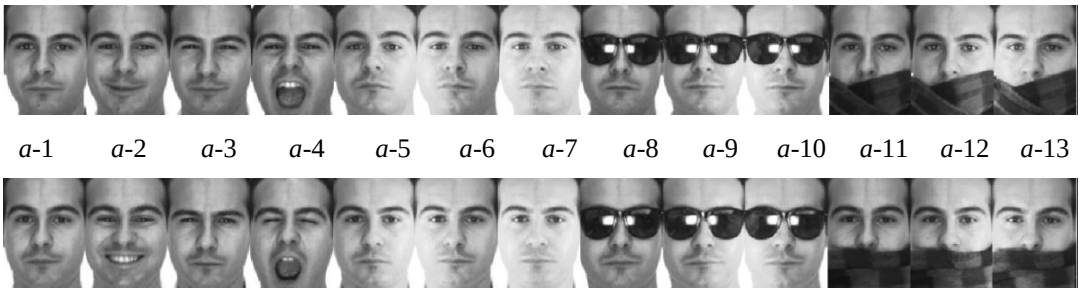


Figure 3.3 AR face database images after the aligning, scaling, localizing, and cropping operations.

3.2.2 ORL face database

The ORL face database contains 40 people with 10 images. The images were taken at different times, under varying light conditions, with different facial expressions and details. The images were taken at homogeneous dark background and all individuals are in frontal position with tolerance to some side movement. The size of each image is 112×92 pixels with 256 gray levels so that 10,304 dimensional feature vectors are obtained from the face images. Images of a subject from ORL face database are shown in Figure 3.4.

Figure 3.4 Images of a subject from ORL face database

3.2.3 YALE face database

YALE database contains 11 images from each of the 15 subjects. Database includes six facial expressions (neutral, happiness, sadness, sleepiness, surprise, and wink) and three illumination conditions (center-light, left-light, and right-light), also subjects wear glasses. The size of the images is 320×243. It is pointed out in [20] that two images of the subjects numbered 2, 3, 6, 7, 8, 9, 12, and 14 are the same. We reduced the database as described in [20] and the number of images from each class is 10. All images are rotated, resized such that the eyes

of the subjects in each image are in the same coordinates. Finally images are cropped and the final size of the images is 120×110. Images of a subject from YALE database and their preprocessed versions are shown in Figure 3.5-(a) and Figure 3.5-(b) respectively.

(a)

(b)

Figure 3.5 Images of a subject from YALE face database. (a) images with their original size, (b) the images after the preprocessing steps.

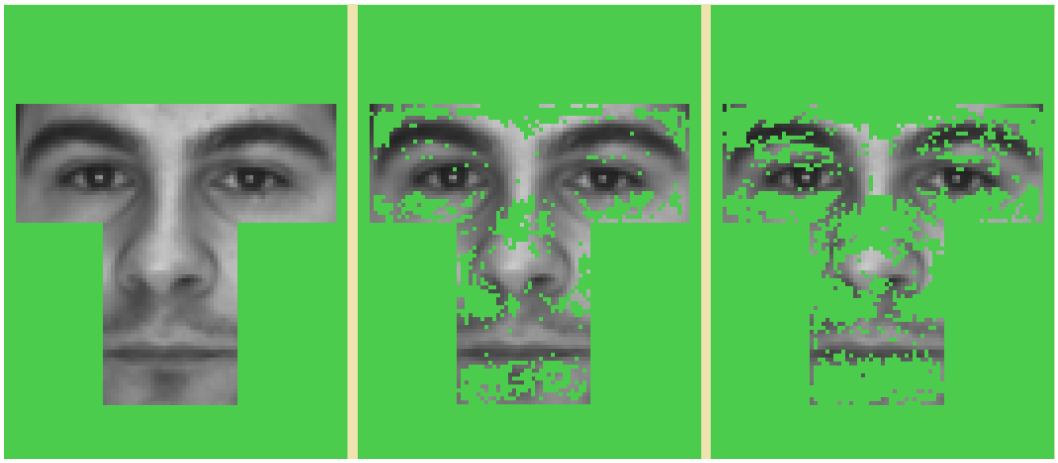
3.2.4 The face images cropped elliptically and in T −shape

It is seen in Figure 3.2 that important parts of the face for recognition purposes are eyes, eyebrows, noses, and lips. However some parts of the hair and

the background behind the left and right side of the neck seem to be also important in the recognition of the face image. One would normally assume that these parts of the face would not be important at all to recognize a person. This situation may be arising due to the face images used in the training sets. Therefore instead of spending much effort to select the pixels in face images in the training set correctly, we cropped the face images into elliptical and T −shaped regions both in the training and test sets of the AR and ORL databases. To have the elliptic mask, we first found an ellipse that passes through the 4 midpoints of all sides of the rectangular image. Then the pixels which are in the interior of the ellipse were automatically selected. Due to the unused exterior pixels, i.e., for AR face database 2004 pixels are eliminated from each face image initially. The way that we formed the T −shaped mask is best illustrated in the first image of Figure 3.7-b. The T −shaped mask is formed so that it covers the most important regions of a face including eyes, nose, and lips. For example, in AR face database we eliminated 6011 pixels by applying this mask. Figure 3.6 shows the two aforementioned masks. Both of these masks are to include only the important parts like eyes, eyebrows, noses, and lips of the face images and eliminate hair and neck areas. The T −shaped mask has an additional elimination of the pixels around the cheek and chin parts of the faces. After obtaining the rectangular, elliptic, and T −shaped face images, we applied our pixel elimination method further. The purpose of these pixel eliminations was to see the change in the recognition rates versus the number of remaining pixels. The face images of AR database cropped elliptically and its variants with 2000, 4000 eliminated pixels are shown in Figure 3.7-a. The same image is cropped in T −shaped regions and its variants with 1000, 2000 eliminated pixels are shown in Figure 3.7-b In the experiments, DCVA, CVA, Eigenface, Fisherface, and LRC stand for the full images, DCVA-E, CVA-E, Eigenface-E, Fisherface-E, and LRC-E stand for the images cropped elliptically, and DCVA-T, CVA-T, Eigenface-T, Fisherface-T, and LRC-T stand for the images cropped in T-shape.

Figure 3.6 The masks used in cropping the images (*a*) elliptically, (*b*) in T −shape.

(a)



(b)

Figure 3.7 A face image from AR face database cropped (*a*) elliptically, (*b*) in T −shape and its variants with eliminated pixels.

3.2.5 Experiments in AR face database

We executed two experiments in this database. In the first experiment fourteen non-occluded images (*a*-1 to *a*-7 and *b*-1 to *b*-7 in Figure 3.3) of the subjects are used from AR database. The 7 images from each class were selected randomly for the training and the remaining images were used in the test stage. Thus 350 images were used in training and 350 images were used in testing. This process is performed 10 times and recognition rates are obtained by averaging each run. 1000 pixels were eliminated according to our pixel selection method and the recognition operations were performed. Pixel elimination was continued until the final dimension of the feature vector is 2005. Recognition rates according to the dimension of the feature vector are given together with the databases of the face images cropped elliptically and in $T-$shape for all methods in Figure 3.8. The recognition rates of the rectangular face images are in general superior to the face images which have elliptical and $T-$shaped regions. When the dimension of the feature vectors is reduced however elliptically cropped images outperform the rectangular and T-shape cropped images. All of the figures belonging to the experiments show the same behavior. There is a slight decrease in all methods with rectangular images until the dimension of the feature vectors is 6000. $T-$shape cropped images always give the worst recognition results. This may be due to the elimination of the discriminative pixels of the face images while using the $T-$shaped mask. In Figure 3.8, recognition results achieved with the elliptically cropped images surpass the recognition results obtained with rectangular images in all methods through the steps of the pixel elimination process. For example, in Figure 3.8-(d) Fisherface-E surpasses Fisherface after the elimination of 5000 pixels.

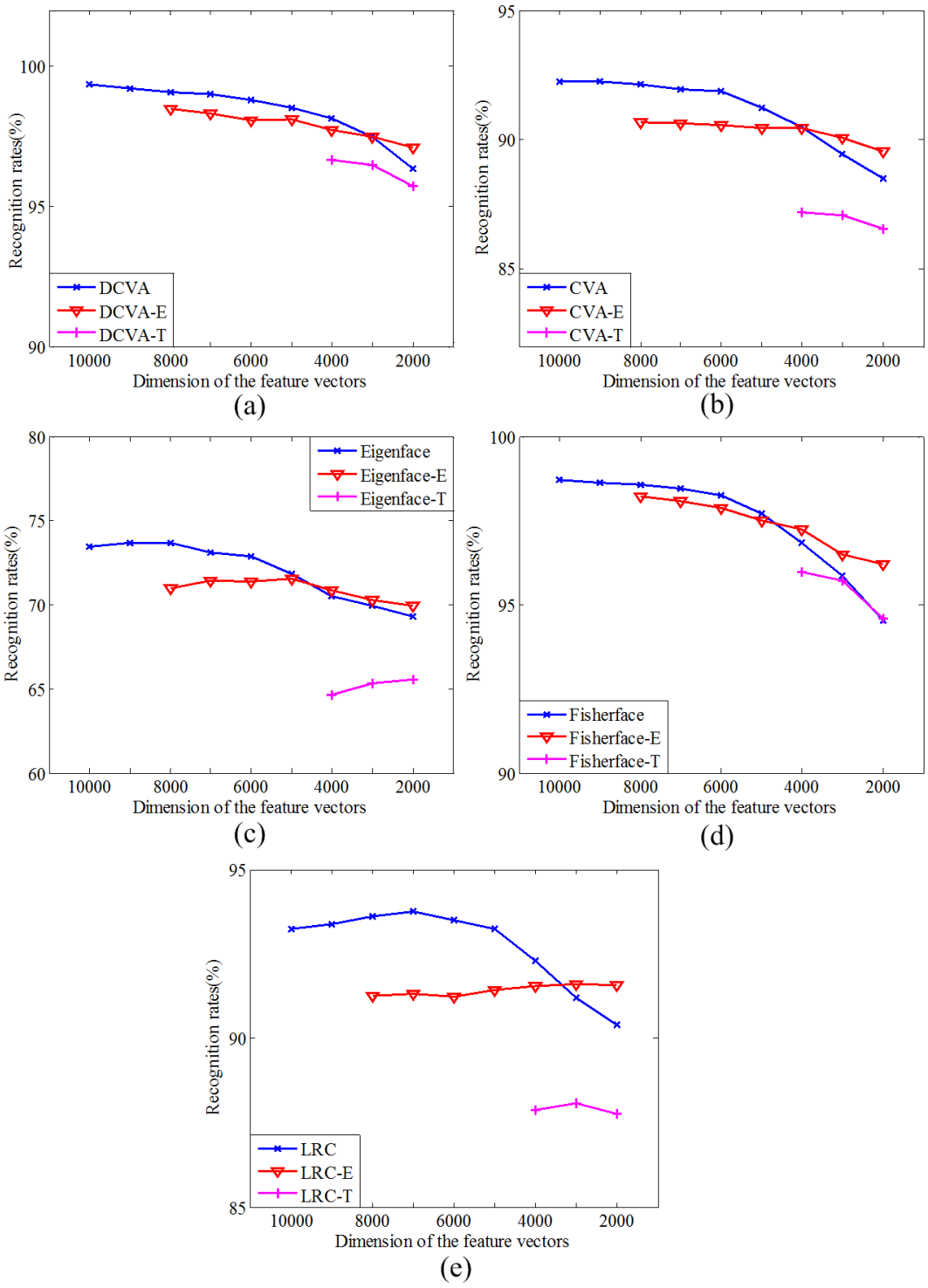
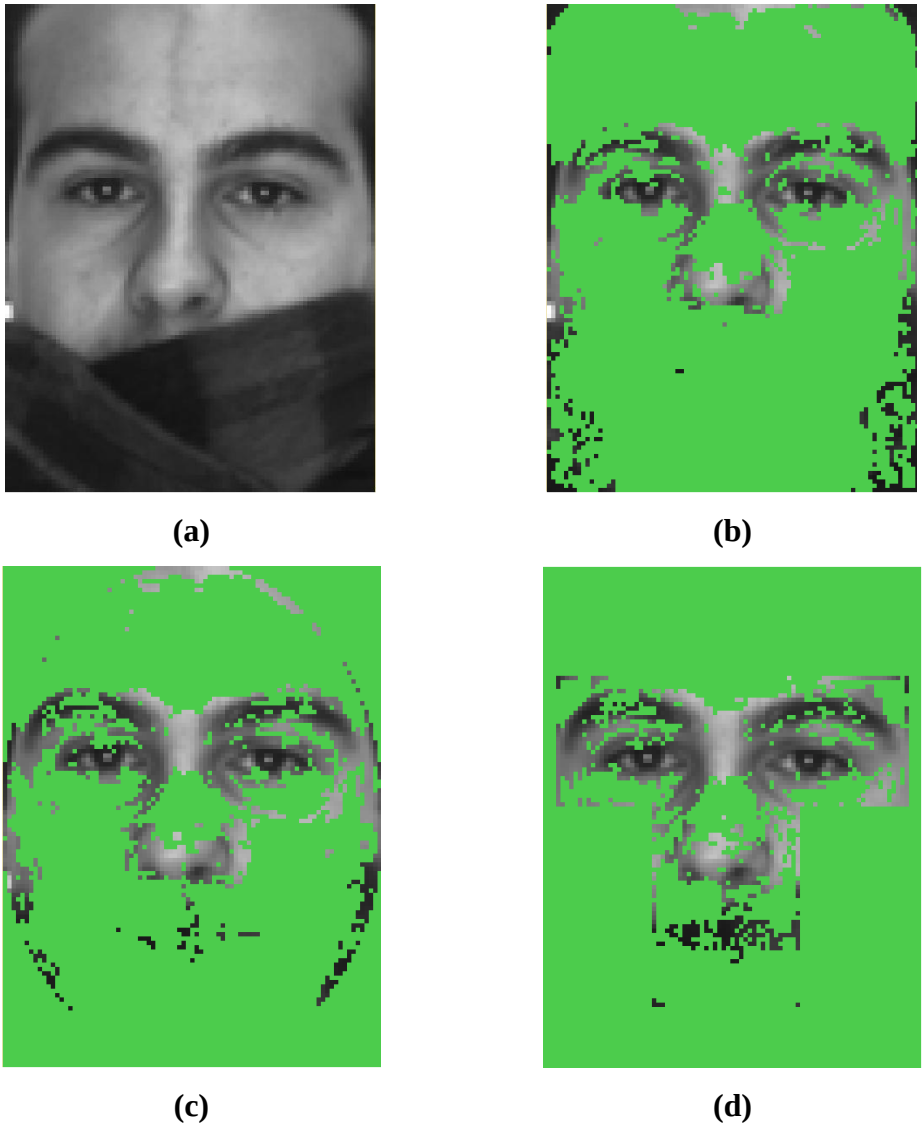


Figure 3.8 Recognition rates of (a) DCVA, (b) CVA, (c) Eigenface, (d) Fisherface, and (e) LRC in AR face database with original images, the face images cropped elliptically and in T −shape with respect to the dimension of the feature vectors.

In the second AR face database experiments, 3 non-occluded (*a*-1, *a*-2, *a*-3 in Figure 3.3) and 3 occluded images (*a*-11, *a*-12, *a*-13 in Figure 3.3) were used for training stage, 3 non-occluded(*b*-1, *b*-2, *b*-3 in Figure 3.3) and 3 occluded images (*b*-11, *b*-12, *b*-13 in Figure 3.3) were used for the test stage to investigate the performance of our feature selection method on the occluded images. In Figure 3.9 the eliminated pixels in an image with a scarf is shown. Especially in Figure 3.9 (b), it is seen that many of the pixels belonging to scarf are eliminated.



(a) **(b)**

(c) **(d)**

Figure 3.9 (a) Original face image and its variants with eliminated pixels: (b) rectangular face image, (c) face image cropped elliptically, (d) face image cropped in T −shape. In (b), (c) and (d) about 8000 pixels are eliminated.

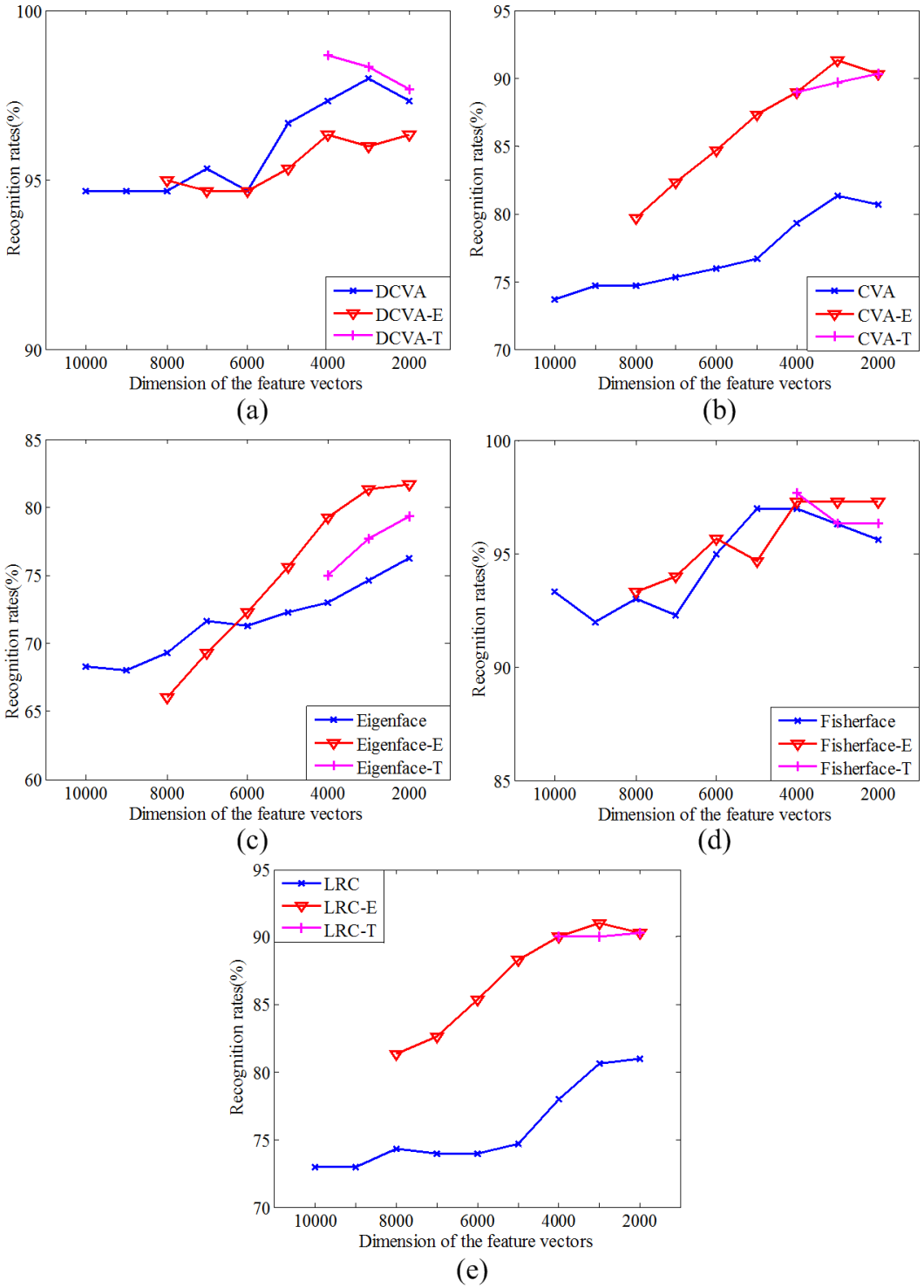


Figure 3.10 Recognition rates of (a) DCVA, (b) CVA, (c) Eigenface, (d) Fisherface, and (e) LRC in AR face database (including occluded images) with original images, the face images cropped elliptically and in T −shape with respect to the dimension of the feature vectors.

The performance of the proposed feature selection method in terms of recognition rates according to the dimension of the feature vector is given with the databases of the face images cropped elliptically and in T −shape regions for all methods as shown in Figure 3.10. It can be seen from the figure that eliminating the pixels using the proposed method increases the recognition rates with almost all types of images and with all methods. In Figure 3.10, generally results with elliptically cropped images are better than the other methods except in Figure 3.10-(a). The recognition rate curves in all figures show an increase as the dimension of the feature vectors decreases in all methods. This result is an important performance indicator of the proposed feature selection method. The best recognition rates of the methods are given in descending order, DCVA, Fisherface, CVA, Eigenface as reported in [20,26].

A great dimensionality reduction is achieved with a small decrease in recognition rates in the first experiments. As an example, if 10,005 dimensional feature vectors are used, the recognition rate is 99.3% using DCVA. However 96.3% recognition rate is achieved by using only 2005 dimensional feature vectors using the same method. Some pixels from background of the original image are selected with the proposed feature selection method which may unfavorably affect the training stage. Since it is known that most discriminatory features are eyes, eyebrows, nose and lips, the face image is cropped to obtain an ellipse and/or a T −shaped region that includes the fiduciary regions of the face image. It can be seen in Figure 3.8 that all methods exhibit the same behavior as the dimension of the feature vectors is decreased. This means that great dimensionality reduction with a proper feature selection method is achieved causing a small recognition rate loss.

In the second part of the experiments in AR face database, the efficiency of the proposed feature selection algorithm becomes clear with the occluded images. It must be noted that establishing the training set has a fundamental role in pixel selection process. In the training set we had to use occluded images to eliminate the occluded regions. Eliminating pixels, this time, not only decreases the testing time but also increases the recognition performance with all methods. As an example, the best recognition result achieved is 98.7% with DCVA-T where the

dimension of the feature vectors is about 4000. The dimension of the feature vectors is 10,005 without pixel elimination and the recognition rate is 94.7%. Thus we achieved 60% dimensional reduction with 4% increase in recognition rate using T −shaped images. The maximum recognition rate with original rectangular shaped face images with 3000 dimensional feature vectors is 98%. Thus a reduction of 70% in feature vector dimension with an increase of 3.3% in recognition rate was achieved. Similar cases occur with the other three methods.

3.2.6 Experiments in ORL face database

In the training stage 5 images per class were randomly selected. 200 images were used in training and the remaining 200 images were used for testing purposes. This procedure repeated 10 times and the recognition rates were obtained by averaging each run. First 1000 pixels were eliminated according to our pixel selection method and the recognition operations described above were performed. Pixel elimination was continued until the final dimension of the feature vector is 2304. Recognition rates are given with respect to the dimension of the feature vector together with the face images cropped elliptically in Figure 3.11. The rectangular shaped face images are always superior to the elliptically cropped face images. This may due to the pose of the subjects in the database. The images are taken in frontal position with tolerance to some side movements as in Figure 3.4. Consistent with the previous experiments, all the graphics in the figure exhibit the same behavior. These experiments show that the proposed feature selection method is sensitive to the side movements of the face.

We cropped the images only in ellipse shape because the faces in ORL database have certain movements. This situation prevents cropping the images in T −shaped regions which would include eyes, eyebrows, mouth and lips at the same time. For instance, if the full image is used as a feature vector, then the recognition rate is 97.5% with DCVA. However the recognition rate is 95% when 4300 dimensional reduced image is used. It is about 58% dimensional reduction with only %2.5 recognition rate loss using DCVA. In this database, rectangular images give the best recognition performance with all four methods. Main reason

is the variable side movements of faces in this database. DCVA is superior to the others in terms of recognition performance in all databases.

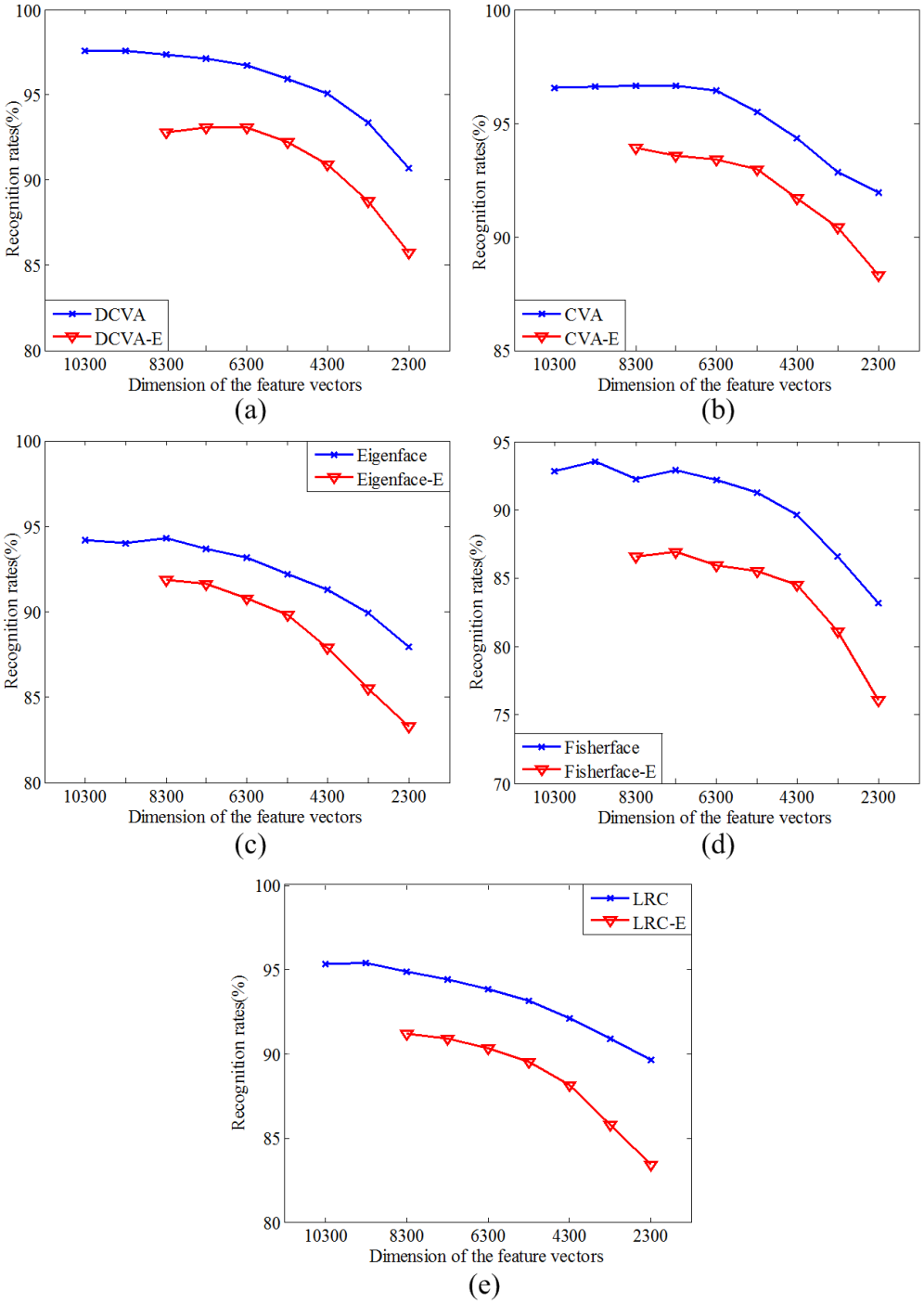


Figure 3.11 Recognition rates of (a) DCVA, (b) CVA, (c) Eigenface, (d) Fisherface, and (e) LRC in ORL face database with original images and the face images cropped elliptically with respect to the dimension of the feature vectors.

3.2.7 Experiments in YALE face database

In the training stage 5 images per class were randomly selected. 75 images were used in training and the remaining 75 images were used for testing purposes. This procedure repeated 10 times and the recognition rates were obtained by averaging each run. First 1000 pixels were eliminated according to our pixel selection method and the recognition operations described above were performed. Pixel elimination was continued until the final dimension of the feature vector is 2200. Recognition rates are given with respect to the dimension of the feature vector together with the face images cropped elliptically and in T −shape for all methods in Figure 3.12. The rectangular shaped face images are generally superior to the face images cropped elliptically and in T −shape. We achieved not only dimensionality reduction but also achieved slight increase in recognition performance with all types of images with all methods, except Fisherface method with rectangular images. In Table 3.1, we summarized the best recognition rate increase and the corresponding dimension reductions of the images as a percentage. Here, Image, Image-E, Image-T, R.I., D.R. stand for the full image, image cropped elliptically, image cropped in T −shape, recognition rate increase, and corresponding dimension reduction respectively.

Table 3.1 Dimensionality reduction amounts as percentages according to the best recognition rates with all methods, using Image, Image-E, and Image-T.

	Image		Image-E		Image-T	
Method	R.I. (%)	D.R. (%)	R.I. (%)	D.R. (%)	R.I. (%)	D.R. (%)
DCVA	0,5	38	0,4	53	1,1	76
CVA	2	53	2,5	45	4,1	83
Eigenface	1,5	38	1,3	76	8	83
Fisherface	0	0	0,4	68	3,8	83
LRC	4,5	45	1,2	76	4	83

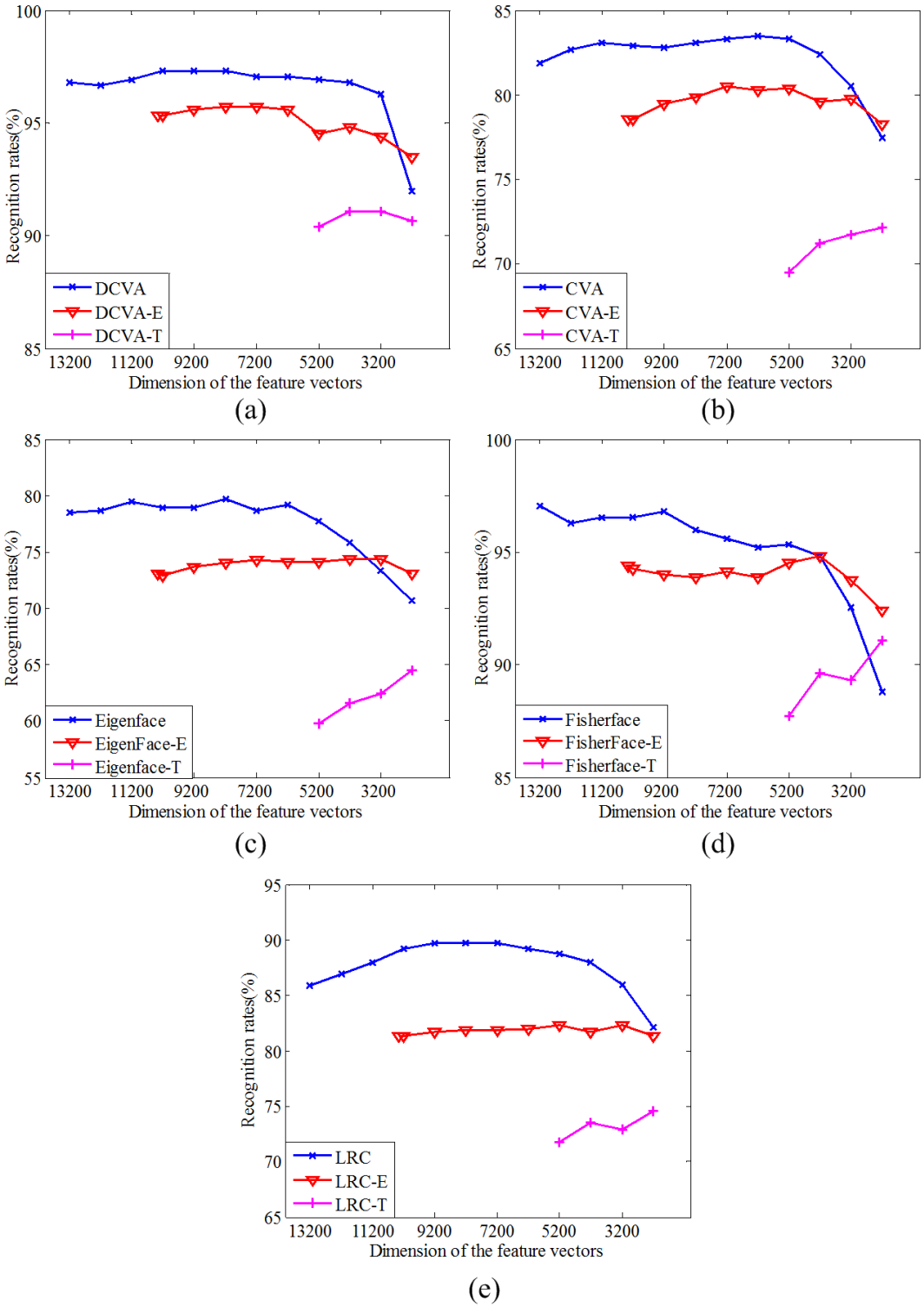


Figure 3.12 Recognition rates of (a) DCVA, (b) CVA, (c) Eigenface, (d) Fisherface, and (e) LRC in YALE face database with original images, the face images cropped elliptically and in T −shape with respect to the dimension of the feature vectors.

It must also be noted that we achieved a great dimensionality reduction with all methods with all types of images with either no or very little loss in recognition rates. For example, using rectangular images with DCVA we achieved 68% dimensional reduction with no recognition loss. Similarly we achieved 83% dimensional reduction only %3 recognition rate loss using elliptically cropped images with CVA. Consistent with the previous experiments, all the graphics in the figure exhibit the same behavior.

3.3 Summary of Pixel Selection

In this chapter, we proposed a novel feature selection method which uses the projection matrix $\boldsymbol{W}$ of the range space of the common vectors. Features are selected according to the norms of columns of $\boldsymbol{W}$. Experiments are performed on AR, ORL, and YALE face databases. Number of dimensions is greatly reduced in AR and ORL databases with and acceptable recognition rate loss in non-occluded images. In the second experiments made with occluded face images of AR face database, not only dimension of the feature vector is reduced but also recognition rate is increased with all four methods. Great dimensionality reduction is achieved with small increase in recognition rates in YALE face database. The results show the success of the proposed feature selection algorithm. Intuitively, it is expected that the important parts of the face for recognition should be the eyes, the mouth, and the nose. In Figure 3.2, Figure 3.7, and Figure 3.9 it is clearly seen that the most important pixels appear around eyes, nose, and mouth. Thus we have used a mathematical approach and we have shown that this intuition is correct experimentally. But one also has to be careful in the selection of faces in the training set, otherwise the results can be delusive. If the face images of a person in the training set have the same background, then the background pixels in the image may be assigned as important features for face recognition as in Figure 3.2.

The testing time is the time required to classify a test image. For example, in AR face database, testing time is reduced from 15.2 milliseconds to 3.2 milliseconds when the dimension of the feature vectors is reduced from 10,005 to 2005, which is important for real-time and real-life applications.

Experimental and theoretical works show that the feature selection method introduced in this thesis is good for dimensionality reduction both in terms of testing time and classification performance.

4 SINGLE IMAGE PER SUBJECT PROBLEM

In this chapter we deal with the one sample problem in face recognition which is a problem for security, law enforcement, person identification, etc. If only one image per person is available, the recognition process gets more difficult. This problem is called *one sample problem* [13]. In the case of having one sample problem, many methods like FLDA which uses the within-class scatter matrix will fail because the within-class scatter matrices are all zero. Traditional methods will suffer or fail when single image per person is available [21,23,45]. Several algorithms have been proposed to overcome this difficulty [13,14,16,17,18,52]. General tendency at these methods is generating the virtual samples to increase the training set size. But this is not the solution of the singularity problem because in face recognition problems dimension of the feature space is high with respect to the number of feature vectors. One solution to overcome the singularity problem is using the two dimensional variant of one dimensional methods after increasing the training set size.

We give two image decomposition methods in this chapter. Singular value decomposition (SVD) based image decomposition method was proposed in [14] to overcome one sample problem. In [15], we developed a novel image feature extraction method using QR decomposition with column pivoting (QRCP) [70,71,72]. Also we propose a two dimensional extension of discriminative common vector approach (2D-DCVA). The performance of 1D-DCVA, 2D-DCVA, and 2D-FLDA are compared in the experimental work.

4.1 Image Decomposition Using QRCP Decomposition

QR decomposition [70,73] is a typical factorization of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. $\boldsymbol{A}$ can be decomposed as $\boldsymbol{A} = \boldsymbol{QR}$ where $\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ with orthogonal columns which span the same subspace with the columns of $\boldsymbol{A}$, and $\boldsymbol{R}$ is an upper triangular matrix.

We decompose the image $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and its transpose $\boldsymbol{A}^T \in \mathbb{R}^{n \times m}$ into two parts using QRCP. QR-decomposition with column pivoting is a modified version

of QR. This algorithm sorts the columns of matrix $\boldsymbol{A}$ such that the absolute values of the diagonal elements of matrix $\boldsymbol{R}$ are in the descending order and this makes a typical energy compaction into some basis images.

Let $\boldsymbol{A} = [\boldsymbol{a}_1 \vdots \boldsymbol{a}_2 \vdots \cdots \vdots \boldsymbol{a}_n] \in \mathbb{R}^{m \times n}$ be a matrix which will be decomposed where $\boldsymbol{a}_i, i = 1, \dots, n$ is the i^{th} column of the image $\boldsymbol{A}$. The algorithm of QRCP-decomposition can be summarized as follows [15,70]:

Table 4.1. QRCP-decomposition algorithm

1. Find the column of $\boldsymbol{A}$ which has the maximum norm, $k = \underset{j}{\operatorname{argmax}}\{\|\boldsymbol{a}_j\|\}$, $j = 1, \dots, n$
2. Swap $\boldsymbol{a}_1$ with $\boldsymbol{a}_k$
3. Compute the 1^{st} column of $\boldsymbol{Q}$, $\boldsymbol{q}_1 = \frac{\boldsymbol{a}_1}{\|\boldsymbol{a}_1\|}$.
4. **for** j **from** 2 **to** n
5. **for** u **from** j **to** n
6. $\boldsymbol{a}_u^* = \boldsymbol{a}_u - \sum_{i=1}^{j-1} \boldsymbol{q}_i \boldsymbol{q}_i^T \boldsymbol{a}_u$
7. **end**
8. $k = \underset{u}{\operatorname{argmax}}\{\|\boldsymbol{a}_u^*\|\}$, $u = j, \dots, n$
9. Swap $\boldsymbol{a}_j$ with $\boldsymbol{a}_k$
10. Compute the j^{th} column of $\boldsymbol{Q}$, $\boldsymbol{q}_j = \frac{\boldsymbol{a}_j^*}{\|\boldsymbol{a}_j^*\|}$.
11. **end**

Then the selection orders of the columns of $\boldsymbol{A}$ are sorted in the permutation matrix $\boldsymbol{P}$. Finally the following equation holds,

$$\boldsymbol{Q}^T \boldsymbol{A} \boldsymbol{P} = \boldsymbol{R} \tag{4.1}$$

Here $\boldsymbol{R}$ is the upper triangular matrix.

After applying the algorithm, the absolute values of the diagonal elements of matrix $\boldsymbol{R}$ are in descending order. The approximation of image $\boldsymbol{A}$ can be written as

$$\widehat{\boldsymbol{A}} = \sum_{i=1}^{k} \boldsymbol{q}_i \boldsymbol{q}_i^T \boldsymbol{A} = \left[\sum_{i=1}^{k} \boldsymbol{q}_i \boldsymbol{\tau}_i \right] \boldsymbol{P}^{-1} \tag{4.2}$$

where $\boldsymbol{q}_i$ is the i^{th} column of $\boldsymbol{Q}$ and $\boldsymbol{\tau}_i$ is the i^{th} row of $\boldsymbol{R}$ and $k \leq m$. The value of k is determined according to the following ratio.

$$\frac{\sum_{i=1}^{k} d_i}{\sum_{i=1}^{m} d_i} \geq E \tag{4.3}$$

where $d_i, i = 1,2, \dots, m$ are the absolute values of the diagonal elements of $\boldsymbol{R}$.

The algorithm given in Table 4.1 concentrates the energy in some basis images $\boldsymbol{q}_i \boldsymbol{\tau}_i$. It can be easily seen in (4.2) that the difference between $\widehat{\boldsymbol{A}}$ and $\boldsymbol{A}$ decreases when k approaches to m.

There are two main directions in a face image, horizontal and vertical. Horizontal and vertical gradients are useful to find the horizontal and vertical edges respectively which are generally seen around boundaries of face, eye, nose, and mouth [51]. Horizontal and vertical variations which most probably occur around the edges [14] contain important information about the within-class scatter. Horizontal and vertical variations of an image can be found by applying QRCP decomposition to the face image A and its transpose A^T respectively. The approximation of A^T which will be called as $\check{A}$ also can be calculated using (4.2). The absolute values of the diagonal elements of R evaluated from QR decomposition and QRCP decomposition are shown in Figure 4.1 and Figure 4.2 respectively.

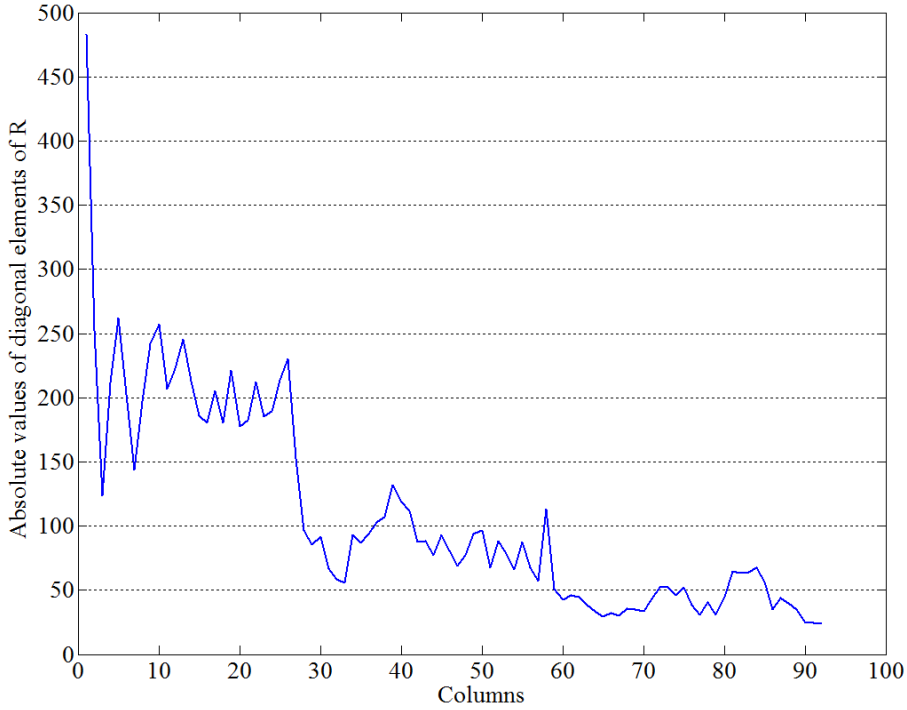


Figure 4.1 The absolute values of diagonal elements of ***R*** evaluated from QR decomposition

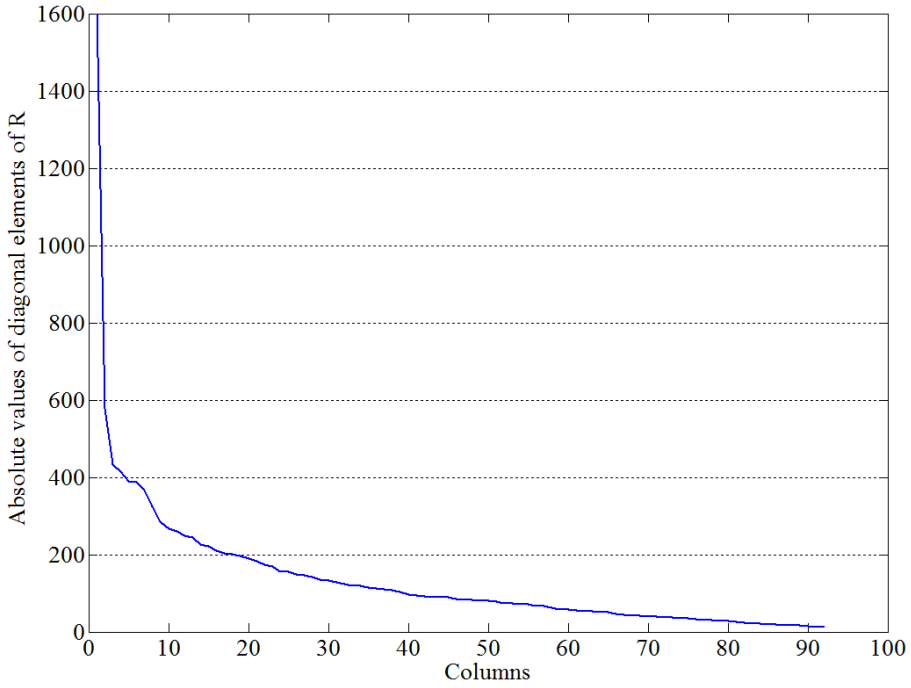


Figure 4.2 The absolute values of diagonal elements of ***R*** evaluated from QRCP decomposition

An image taken from a subject from the ORL face database and two virtual images $\hat{A}, \check{A}$ reconstructed from the image and from the transpose of the image are shown in Figure 4.3. The within-class covariance matrices in (2.18) and in (2.25) can be calculated using the generated set $\{A, \hat{A}, \check{A}\}$. Using this set, not only FLDA is made applicable but also the training set size is increased to three. It is known that increasing the training set size helps us in modeling the classes better and this will increase the performance of the recognition system [7,11]. Let us define the difference images as $\varepsilon_1 = A - \hat{A}, \varepsilon_2 = A - \check{A}, \varepsilon_3 = \hat{A} - \check{A}$. Using the difference images $\boldsymbol{S}_W$ can be calculated as below:

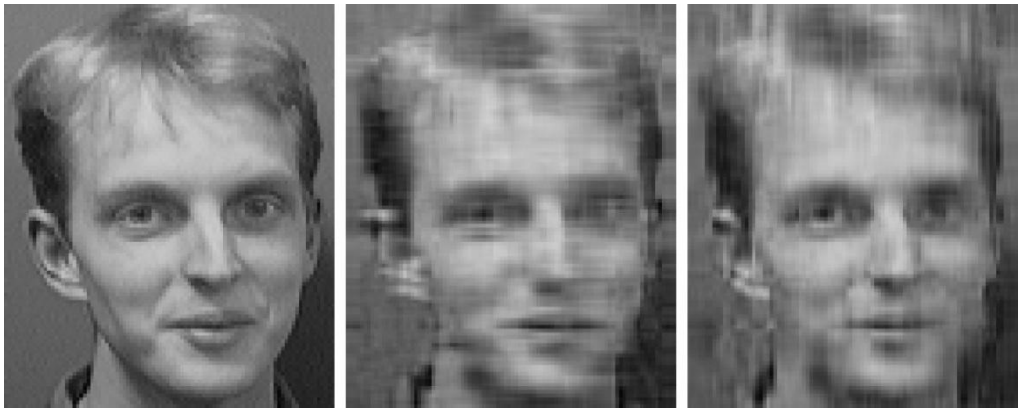
$$S_W = \sum_{i=1}^{C} \sum_{j=1}^{N} \left(A_j^i - M^i\right)^T \left(A_j^i - M^i\right) \tag{4.4}$$

$$= \sum_{i=1}^{C} \left[\left(A^i - \frac{(A^i + \hat{A}^i + \check{A}^i)}{3}\right)^T \left(A^i - \frac{(A^i + \hat{A}^i + \check{A}^i)}{3}\right) + \left(\hat{A}^i - \frac{(A^i + \hat{A}^i + \check{A}^i)}{3}\right)^T \left(\hat{A}^i - \frac{(A^i + \hat{A}^i + \check{A}^i)}{3}\right) + \left(\check{A}^i - \frac{(A^i + \hat{A}^i + \check{A}^i)}{3}\right)^T \left(\check{A}^i - \frac{(A^i + \hat{A}^i + \check{A}^i)}{3}\right) \right] \tag{4.5}$$

$$= \frac{1}{3} \sum_{i=1}^{C} \left[\left(A^i - \hat{A}^i\right)^T \left(A^i - \hat{A}^i\right) + \left(A^i - \check{A}^i\right)^T \left(A^i - \check{A}^i\right) + \left(\hat{A}^i - \check{A}^i\right)^T \left(\hat{A}^i - \check{A}^i\right) \right] \tag{4.6}$$

$$= \frac{1}{3} \sum_{i=1}^{C} \sum_{j=1}^{3} \left(\varepsilon_j^i\right)^T \varepsilon_j^i \tag{4.7}$$

This means that S_W can be modeled using the difference images which contain horizontal and vertical variations. The difference images are shown in Figure 4.4. A typical system diagram of the QRCP-based recognition system is shown in Figure 4.5.



(a) **(b)** **(c)**

Figure 4.3 (a) The original image, approximated images evaluated (b) from the original image $\hat{A}$ and (c) from the transpose of the original image $\check{A}$.

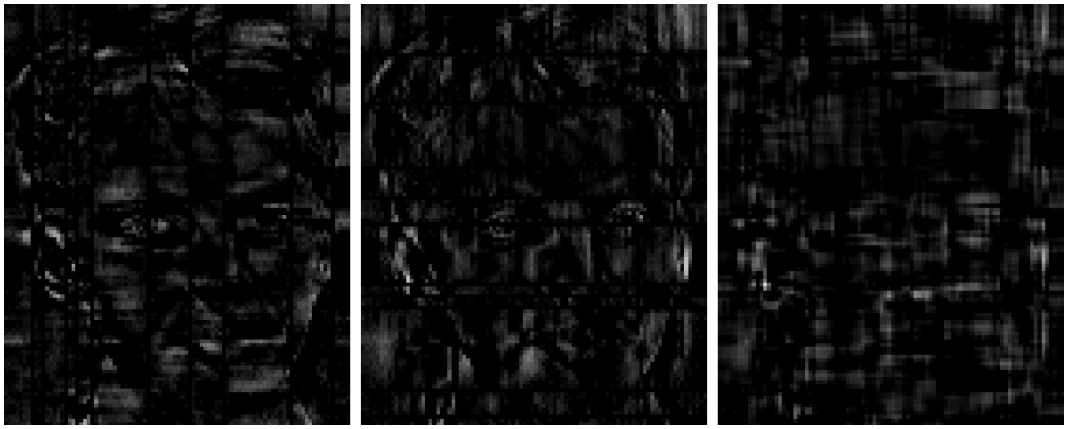


Figure 4.4 The difference images $\varepsilon_1, \varepsilon_2, \varepsilon_3$ respectivelly.

4.2 Image Decomposition Using SVD Decomposition

Let $\boldsymbol{A}$ be an $m \times n$ dimensional image. $\boldsymbol{A}$ can be decomposed using singular value decomposition as follow [74].

$$\boldsymbol{A} = \boldsymbol{U} \begin{bmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^T \tag{4.8}$$

where $\boldsymbol{U}_{m\times m}$, $\boldsymbol{V}_{n\times n}$ are orthogonal matrices and $\boldsymbol{D}_{r\times r}$is the diagonal matrix whose diagonal elements are the singular values $\sigma_i, i = 1, \dots, r$. It is shown in [14] that most of the energy of the image $\boldsymbol{A}$ is concentrated in the basis images

$$\boldsymbol{I}_i = \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T, i = 1, \dots, n \tag{4.9}$$

corresponding to the largest singular values. Here $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ represent the i^{th} columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. Assuming that $p < n$ the approximated image evaluated from SVD can be calculated as follow.

$$\overline{\boldsymbol{A}} = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{4.10}$$

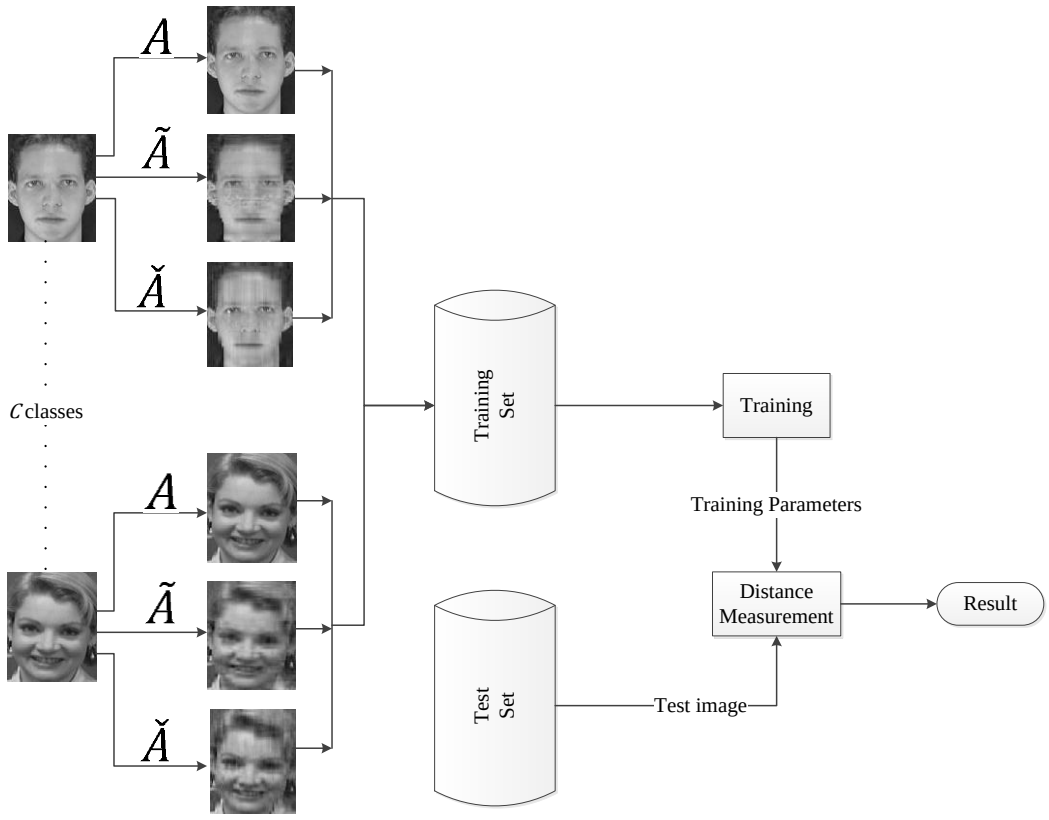


Figure 4.5 System diagram [15].

In Figure 4.6 an image from ORL database and the virtual image generated from the image using SVD are shown. We can compute the within-class covariance matrices in (2.18) and in (2.25) using the generated set $\{\boldsymbol{A}, \overline{\boldsymbol{A}}\}$.

(a) **(b)**

Figure 4.6 (a) The original image, (b) the approximated image evaluated using SVD.

4.3 Two Dimensional Extension of Discriminative Common Vector Approach

The face image is transformed from matrix to vector form in many traditional pattern recognition methods [20,21,23,31]. Recently there appear methods which try to extract features without transform the image into vector form [27,53,75,76]. It is also known that generally two dimensional methods outperform their one dimensional variants [66]. In this section we give a two dimensional extension of DCVA.

Let C be the number of image classes, N, be the number of feature vectors in each class and, $\boldsymbol{A}_m^i$ be the m^{th} two dimensional p by q pixel image of the i^{th} class. We convert the image matrix $\boldsymbol{A}_m^i$ to a vector $\boldsymbol{a}_m^i$ in the $n = p \times q$ dimensional space.

It is proved in [20] that the common vectors obtained from total within-class scatter matrix are unique for each class. In order to get unique common

vectors we use $\boldsymbol{\Phi}_T$ which is defined in (2.18), in the first stage of the proposed method. We apply the eigen decomposition to $\boldsymbol{\Phi}_T$ and obtain the projection matrix $\boldsymbol{U} = [\boldsymbol{u}_1 \boldsymbol{u}_2 \dots \boldsymbol{u}_{NC-C}]$ where $\boldsymbol{u}_i, i = 1,2, \dots NC - C$ are the eigenvectors corresponding to the nonzero eigenvalues of $\boldsymbol{\Phi}_T$. Then the common vector of i^{th}class is calculated as

$$\boldsymbol{a}_{com}^{i} = \boldsymbol{a}_{m}^{i} - \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{a}_{m}^{i}, \;\; m = 1, \dots, N, \;\; i = 1, \dots, C \tag{4.11}$$

It should be noted that (2.20) and (4.11) give exactly the same results. We convert the common vectors $\boldsymbol{a}_{com}^{i}$ into p by q matrices, $\boldsymbol{A}_{com}^{i}$. The covariance matrix of the common matrices can be calculated as

$$\boldsymbol{S}_{com} = \sum_{i=1}^{C} \left(\boldsymbol{A}_{com}^{i} - \boldsymbol{A}_{ave}\right)^T \left(\boldsymbol{A}_{com}^{i} - \boldsymbol{A}_{ave}\right) \tag{4.12}$$

where $\boldsymbol{A}_{ave} = 1/C \sum_{i=1}^{C} \boldsymbol{A}_{com}^{i}$ is the mean of the common matrices. We are trying to find the optimal projection vectors $\boldsymbol{W} = [\boldsymbol{w}_1 \vdots \boldsymbol{w}_2 \vdots \cdots \vdots \boldsymbol{w}_d]$ which maximize the criterion $J(\boldsymbol{W}) = \boldsymbol{W}^T\boldsymbol{S}_{com}\boldsymbol{W}$ under the constraint $\boldsymbol{W}^T\boldsymbol{W} = \mathbf{I}$. Here d can be at most $\min(C - 1, n)$.

We use the nearest neighbor classifier for classification. The decision rule is the same as the rule given in Section 2.3.1.

4.4 Experimental Work

In this section we compare the performances of 1D-DCVA, 2D-DCVA, and 2D-FLDA in one sample problem in five different databases, ORL, FERET [77], YALE, UMIST [78], PolyU-NIR [79]. We used the virtual samples generated using both SVD based image decomposition method and QRCP based image decomposition method in the experiments.

4.4.1 FERET face database

The FERET database contains 14,051 images from 1199 different subjects with different illumination conditions, pose, ethnicity, age, and expression. In the experiments we use 200 images from f_a and f_b probes. The original size of the images is 384×256. All images are scaled, aligned, cropped, and resized. The final size of the images is 100×100. Sample images from f_a and f_b probes of FERET database in their original size and after the preprocessing steps are shown in Figure 4.7-(a) and Figure 4.7-(b) respectively.

(a)

(b)

Figure 4.7 Sample images from FERET face database. (a) images with their original size, (b) the images after the preprocessing steps.

4.4.2 UMIST face database

UMIST face database contains 564 images from 20 subjects. Number of images per subject varies from 19 to 48. So we used 19 images from each subject. The size of the cropped images are 112×92 with 256-bit gray-scale. Subjects cover a range of poses from profile to frontal views and gender. In Figure 4.8 the selected images from UMIST database are shown.

Figure 4.8 Images of a subject from UMIST face database.

4.4.3 PolyU-NIR face database

PolyU – NIR face database contains 35,000 images which are captured near infrared band from 350 subjects including different pose, scale, illumination, expression, time, blurring, etc. The original size of each image is 576×768. We selected 420 images from 60 subjects. Each image is normalized according to the location of eyes so that the distance between eyes are the same. Then the image is cropped and the final size of the image is 120×90. Images of three different subjects from PolyU – NIR face database are shown in Figure 4.9.

Figure 4.9 Images of from PolyU – NIR face database.

4.4.4 Experiments

We tested the performances of the methods in single image problem. We used SVD based image decomposition and QRCP based image decomposition methods to generate virtual samples. In equation (4.3), we selected the value of $E = 97\%$ as in [15] and in (4.10) we selected $p = 3$ as in [14]. The number of subjects, the number of images taken from each subject, and the size of the images taken from ORL, FERET, YALE, UMIST, and PolyU-NIR databases after the preprocessing operations are summarized in Table 4.2. In the experiments we select a random image from each class and use that image to generate the virtual sample(s) using SDV based image decomposition or QRCP based image decomposition method. The image and the generated virtual samples are used to form the training set of the class where the image is selected. The rest of the images are used for testing purposes. This procedure is repeated five times and the recognition rates are obtained by averaging each run. We did this procedure to all databases. The recognition rates of the methods with respect to the number of the projection vectors are shown in Figure 4.10.

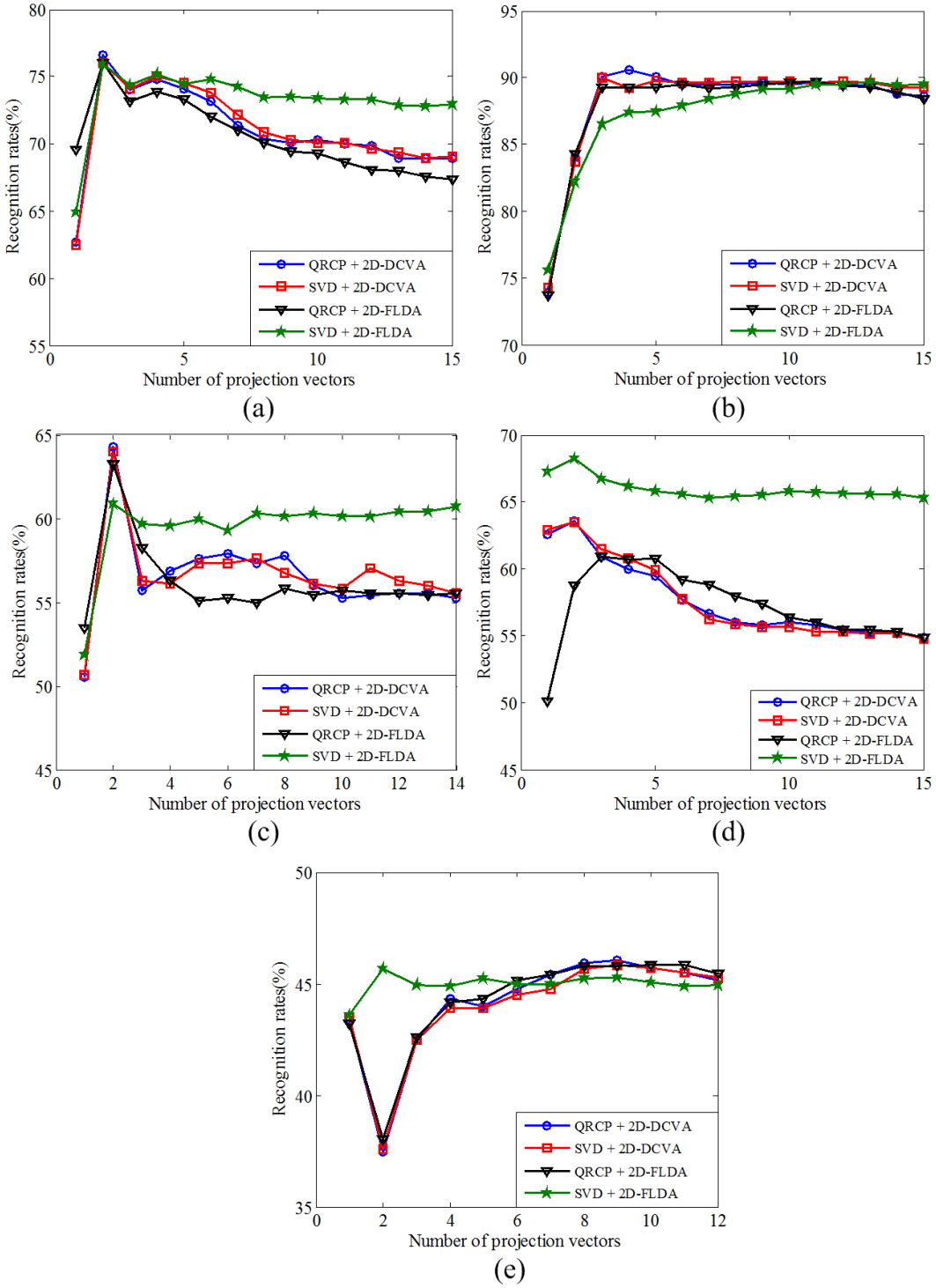


Figure 4.10 The recognition rates of 2D-DCVA and 2D-FLDA using QRCP and SVD based decomposition methods in (a) ORL, (b) FERET, (c) YALE, (d) UMIST, and (e) PolyU-NIR face databases

Table 4.2 The summary of the databases after the preprocessing steps

Database	Number of classes	Number of images per class	Dimensions
ORL	40	10	112x92
FERET	200	2	100x100
YALE	15	10	120x110
UMIST	20	19	112x92
PolyU-NIR	60	7	120x90

Table 4.3 The recognition rates of 1D-DCVA, 2D-DCVA, and 2D-FLDA using SVD based image decomposition

Method	ORL	FERET	YALE	UMIST	PolyU – NIR
1D-DCVA	70.9	87	55.9	58.7	39.7
2D-DCVA	**76.1**	**90**	**64.0**	63.5	**45.8**
2D-FLDA	75.9	89.7	63.3	**68.2**	45.7

Table 4.4 The recognition rates of 1D-DCVA, 2D-DCVA, and 2D-FLDA using QRCP based image decomposition

Method	ORL	FERET	YALE	UMIST	PolyU – NIR
1D-DCVA	69.8	88.8	56.3	55.9	41.1
2D-DCVA	**76.6**	**90.6**	**64.3**	**63.6**	**46.1**
2D-FLDA	76.0	89.7	60.9	60.9	45.8

Table 4.3 and Figure 4.4 show the best recognition rates of 1D-DCVA, 2D-DCVA, and 2D-FLDA methods using SVD based image decomposition method and QRCP based image decomposition method respectively. We can see from Table 4.3 and Figure 4.4 that 2D-DCVA method outperforms 1D-DCVA and 2D-

FLDA methods. In Table 4.3, 2D-FLDA exhibits better performance than 2D-DCVA only in the UMIST face database. 1D-DCVA shows the lowest performance in all experiments. This should be due to fact that the two dimensional methods generally outperform the one dimensional methods [66].

4.5 Summary of Single Image Training Problem

In this chapter, we gave two image decomposition methods, namely SVD based image decomposition and QRCP based image decomposition which are used to generate virtual samples. QRCP which is a modification of QR method decomposes the image into two matrices Q and R. Then the absolute values of the diagonal elements of R are in descending order. By this way most of the energy is concentrated into some basis images. Using some of this basis images an approximation of the original image is generated. Another approximation of the image is generated using the transpose of the image. The image and its two approximations are used to generate the training set of the class where the image belongs to. Similarly SVD is another matrix decomposition method which is first used in one sample problem in [14]. Image is decomposed into some basis images using singular values. An approximation of the image is generated using the basis images corresponding to the largest singular values. Similarly the original image and the approximation of that image are used to form the training set of the subject.

Also, we proposed a novel two dimensional extension of discriminative common vector approach (2D-DCVA). We used the vector form of the images in the first stage of the method to compute the common vectors uniquely.

The performances of 1D-DCVA, 2D-DCVA, and 2D-FLDA are compared in one sample problem in five different databases. It is clearly seen in Table 4.3 and Figure 4.4 that 2D-DCVA surpasses the other two methods. Also we can infer from Table 4.3 and Figure 4.4 that QRCP is a better decomposition method than SVD based image decomposition.

5 COVARIANCE MATRIX ESTIMATION IN SUBSPACES

It is known that the data distribution of a class can be represented by a Gauss distribution. But the variances of a Gauss distribution depend only on the data of the class that the distribution belongs to. In high dimensional data, the equipotential curves of a class distribution become elliptic hypercylinders with endless top and bottom surfaces because of the zero eigenvalues of the covariance matrix. It should be better to use the data of the other classes to model a class that may help to limit the Gauss distribution and convert the equipotential curves of a class distribution to hyperellipsoids. In this work, when modeling the covariance matrix of a class we use not only the data of one class but also the data of the other classes. We try to model the classes in the subspaces, especially in the range space of the total within-class scatter matrix, $\boldsymbol{\Phi}_T$, ($R(\boldsymbol{\Phi}_T)$) in this section.

In face recognition problems, if the whole face image is used as a feature vector, the dimension of the feature space is very high. Since the images from a class are very limited, the class model cannot be estimated well. Let C, M, n be the number of classes, number of feature vectors from each class and the dimension of the feature vectors respectively. We perform a great dimensionality reduction by projecting the feature vectors onto $R(\boldsymbol{\Phi}_T)$. The dimension of the range space of $\boldsymbol{\Phi}_T$ is $C(M-1)$ whereas the dimension of range space of within-class covariance matrix $\boldsymbol{\Phi}_j$ ($R(\boldsymbol{\Phi}_i)$) which is given in (2.22) is $(M-1)$. If we try to model a class in the range space of $\boldsymbol{\Phi}_T$, we do not have enough samples in that class.

When we model a class in this study, we use the data of all classes. We give two methods to model the classes with exponential surfaces, thus we estimate the within-class covariance matrices of the classes in $R(\boldsymbol{\Phi}_T)$. We obtain a decision surface using the covariance matrices of the classes in $R(\boldsymbol{\Phi}_T)$. We illustrate the two methods with numerical examples and finally we compare the performances of two methods with several experiments in YALE, ORL, and AR face database.

5.1 Class Modeling Using Exponential Hypersurfaces

Let C be the number of classes and M be the number of feature vectors from each class in n-dimensional subspace. Let with $C = 2$ and $M = 2$. The training set of the two classes is $C_1 = \{\boldsymbol{x}_1^1,\ \boldsymbol{x}_2^1\}$, $C_2 = \{\boldsymbol{x}_1^2,\ \boldsymbol{x}_2^2\}$ where $\boldsymbol{x}_i^c \in \mathbb{R}^n$. Let $\boldsymbol{W}$ be the projection matrix onto the $R(\boldsymbol{\Phi}_T)$. The projections of the feature vectors onto the $R(\boldsymbol{\Phi}_T)$ are defined as

$$\boldsymbol{y}_i^j = \boldsymbol{W}^T \boldsymbol{x}_i^j, \qquad i,j = 1,2. \tag{5.1}$$

Similarly, the average vectors of two classes can be defined as $\boldsymbol{y}_{ave}^1 = \boldsymbol{W}^T \boldsymbol{x}_{ave}^1$, $\boldsymbol{y}_{ave}^2 = \boldsymbol{W}^T\ \boldsymbol{x}_{ave}^2$.

We try to find the surfaces for classes using the training set data in $C(M-1)+1$-dimensional subspace. Let $z_c = e^{-K^c}$, $c = 1,2$ be the surfaces that will be fitted to the class distributions. Here

$$K^c = (\boldsymbol{y} - \boldsymbol{y}_{ave}^c)^T \boldsymbol{R}_c (\boldsymbol{y} - \boldsymbol{y}_{ave}^c), \qquad c = 1,2 \tag{5.2}$$

where $\boldsymbol{R}_c = \begin{bmatrix} r_1^c & 0 \\ 0 & r_2^c \end{bmatrix}$, $c = 1,2$. We select $\boldsymbol{R}_c$ as diagonal matrices to reduce the number of unknown parameters. As a result, the level curves of the exponential functions become the ellipses which are parallel to the coordinate axes. It can be thought that $\boldsymbol{R}_c$ is the inverse of the estimated covariance matrix of the c^{th} class. Actually, we model the covariance matrix of each class in the range space of $\boldsymbol{\Phi}_T$.

Let $\boldsymbol{y}_i^c$ be the projected feature vector from one of the two classes onto the $R(\boldsymbol{\Phi}_T)$. We use the following conditions to compute the *sum of square errors* for z_1.

$$\boldsymbol{y}_i^c \in C_1 \Rightarrow 1 = e^{-K_{i,1}^1} + \varepsilon_i^1, i = 1,2 \tag{5.3}$$

$$\boldsymbol{y}_i^c \in C_2 \Rightarrow 0 = e^{-K_{i,2}^1} + \varepsilon_i^2, i = 1,2 \tag{5.4}$$

The sum of squares for z_1 is computed as follows

$$\begin{aligned} \mathrm{SES}_1 &= \sum_{c=1}^{2}\sum_{i=1}^{2}(\varepsilon_i^c)^2 \\ &= \left(1-e^{-K_{1,1}^1}\right)^2 + \left(1-e^{-K_{2,1}^1}\right)^2 + \left(e^{-K_{1,2}^1}\right)^2 + \left(e^{-K_{2,2}^1}\right)^2 \end{aligned} \tag{5.5}$$

Similarly, the sum of square errors for z_2 can be computed using the following conditions.

$$\boldsymbol{y}_i^c \in C_1 \Rightarrow 0 = e^{-K_{i,1}^2} + \varepsilon_i^1, i = 1,2 \tag{5.6}$$

$$\boldsymbol{y}_i^c \in C_2 \Rightarrow -1 = e^{-K_{i,2}^2} + \varepsilon_i^2, i = 1,2 \tag{5.7}$$

$$\begin{aligned} \mathrm{SES}_2 &= \sum_{c=1}^{2}\sum_{i=1}^{2}(\varepsilon_i^c)^2 \\ &= \left(1-e^{-K_{1,1}^2}\right)^2 + \left(1-e^{-K_{2,1}^2}\right)^2 + \left(e^{-K_{1,2}^2}\right)^2 + \left(e^{-K_{2,2}^2}\right)^2 \end{aligned} \tag{5.8}$$

The values of $r_1^c, r_2^c, i = 1,2$ that minimize SES_1 and SES_2 are used to obtain the difference surface $z = z_1 - z_2$. The optimum values of $r_1^c, r_2^c, i = 1,2$ are found by *steepest descent method* [80]. In this method, we start the searching at an initial point $\boldsymbol{x}_0$ which is generally chosen randomly. In each step the point is updated according to the following equation.

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \mu \nabla f(\boldsymbol{x}_0) \tag{5.9}$$

Here ∇f denotes the gradient of the function f that will be minimized and μ denotes the weight factor. μ is generally chosen as $0 < \mu < 1$ and it determines the step size at every iteration.

When we model a class we use data of the all classes. If we assume that $\boldsymbol{R}$ is a non-diagonal matrix, the number of unknown variables will increase drastically. For example, the number of unknown variables in two-dimensional

space is 3 whereas there are 1275 unknown variables in 50-dimensional space. The method is not feasible for high dimensional data even if we select $\boldsymbol{R}$ as diagonal matrix.

In classification phase, a query is sent to all two-class models and is assigned to one of two classes in each model. The class with the highest vote is the final predicted class. The assignment is done according to the value of the difference surface $z = z_1 - z_2$. If $z > 0$, then the query belongs to the class 1 else it belongs to the class 2.

5.2 Novel Covariance Matrix Modeling in the Subspaces

When we estimate the new within-class covariance matrix of a class in the range space of $\boldsymbol{\Phi}_T$, we use not only the class's own data but also use the data of the other class for a two-class problem. Let C_1 and C_2 be the two classes that will be modeled. Let $\boldsymbol{x}_i^1$, $\boldsymbol{x}_i^2, i = 1, \dots, M$ be feature vectors and let $\boldsymbol{x}_{ave}^1$, $\boldsymbol{x}_{ave}^2$ be the means of the classes C_1 and C_2 respectively. Let $\boldsymbol{W}$ be the projection matrix of the range space of $\boldsymbol{\Phi}_T$. Then the projections of the feature vectors can be defined as

$$\begin{matrix} \boldsymbol{y}_i^1 = \boldsymbol{W}^T \boldsymbol{x}_i^1 \\ \\ \boldsymbol{y}_i^2 = \boldsymbol{W}^T \boldsymbol{x}_i^2 \end{matrix}, \; i = 1, \dots, M \tag{5.10}$$

and the average vectors in the range space of $\boldsymbol{\Phi}_T$ can be defined as $\boldsymbol{y}_{ave}^1 = \boldsymbol{W}^T \boldsymbol{x}_{ave}^1, \boldsymbol{y}_{ave}^2 = \boldsymbol{W}^T \boldsymbol{x}_{ave}^2$.

The within-class covariance matrices of the classes C_1 and C_2 in the range space of $\boldsymbol{\Phi}_T$ are defined as follow.

$$\boldsymbol{S}_1 = \sum_{i=1}^{M} (\boldsymbol{y}_i^1 - \boldsymbol{y}_{ave}^1)^T (\boldsymbol{y}_i^1 - \boldsymbol{y}_{ave}^1) \tag{5.11}$$

$$\boldsymbol{S}_2 = \sum_{i=1}^{M} (\boldsymbol{y}_i^2 - \boldsymbol{y}_{ave}^2)^T (\boldsymbol{y}_i^2 - \boldsymbol{y}_{ave}^2) \tag{5.12}$$

We also define new covariance matrices that can be used to model the deviation of a data from the mean of the other class.

$$\boldsymbol{Q}_1 = \sum_{i=1}^{M} (\boldsymbol{y}_i^2 - \boldsymbol{y}_{ave}^1)^T (\boldsymbol{y}_i^2 - \boldsymbol{y}_{ave}^1) \tag{5.13}$$

$$\boldsymbol{Q}_2 = \sum_{i=1}^{M} (\boldsymbol{y}_i^1 - \boldsymbol{y}_{ave}^2)^T (\boldsymbol{y}_i^1 - \boldsymbol{y}_{ave}^2) \tag{5.14}$$

The new covariance matrices can be defined as

$$\boldsymbol{\Phi}_1 = \boldsymbol{S}_1 + \frac{1}{M}\boldsymbol{Q}_1 \tag{5.15}$$

$$\boldsymbol{\Phi}_2 = \boldsymbol{S}_2 + \frac{1}{M}\boldsymbol{Q}_2 \tag{5.16}$$

$\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ matrices are normalized with the number of feature vectors, M, used in training set from a class to reduce the unwanted effect in recognition. In a two dimensional space if $M = 2$ the rank of $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are 1. Thus $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are singular matrices. But the covariance matrices defined in (5.14) and (5.15) are nonsingular and invertible.

The new exponential surfaces, z_1 and z_2 can be fitted by redefining $\boldsymbol{R}_c$ in (5.2) as $\boldsymbol{R}_c = \boldsymbol{\Phi}_c^{-1}, c = 1{,}2$ as follow:

$$z_c = \exp(K^c) = \exp\left(-(\boldsymbol{y} - \boldsymbol{y}_{ave}^c)^T \boldsymbol{\Phi}_c^{-1} (\boldsymbol{y} - \boldsymbol{y}_{ave}^c)\right), c = 1{,}2. \tag{5.17}$$

5.3 Numerical Examples

In this section, we give two numerical examples to illustrate the proposed class modeling methods.

Example 5.1 Let $C_1 = \left\{\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}\right\}$, $C_2 = \left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}2\\1\end{bmatrix}\right\}$ are the two classes. In this example we model the classes in the range space of $\mathbf{\Phi}_T$ using the method given in Section 5.1.

After applying the steepest descent algorithm to SES_1 and SES_2, the optimum values of $r_1^c, r_2^c, i = 1,2$ are found as below.

$$[r_1^1 \quad r_2^1 \quad r_1^2 \quad r_2^2] = [2.5073 \quad 0.8753 \quad 0 \quad 1.2797]$$

The mesh and the contour plots of the difference surface are shown in Figure 5.1 and Figure 5.2 respectively. In Figure 5.2, the curve represented by "$-+-$" indicates the zero-crossing level of the difference surface which can be used as decision curve.

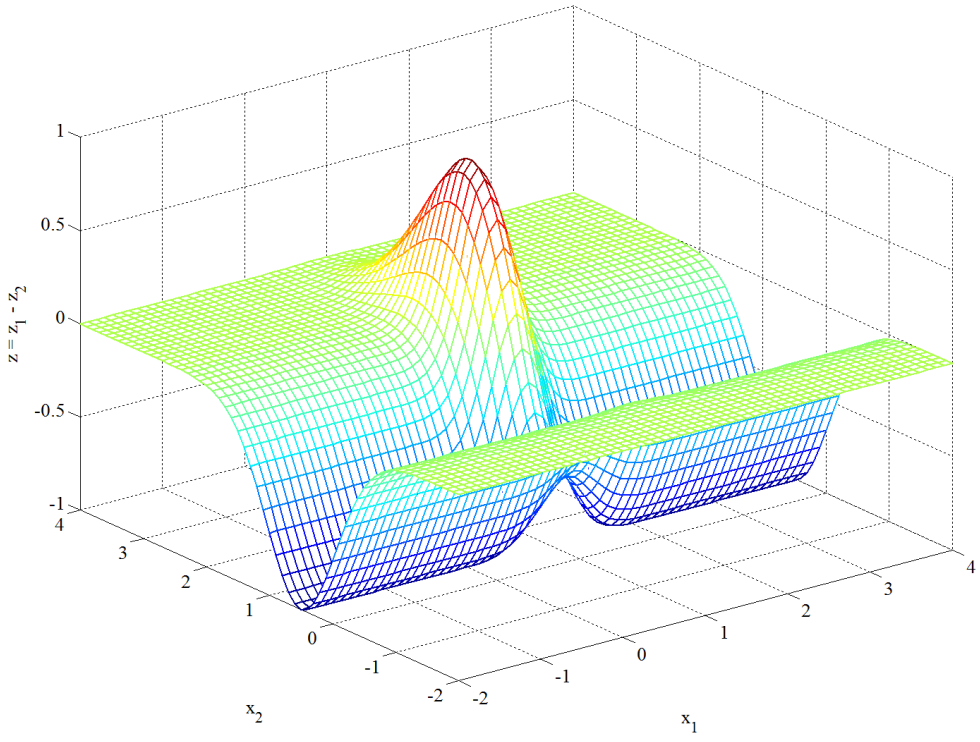


Figure 5.1 The mesh plot of the difference surface $z = z_1 - z_2$

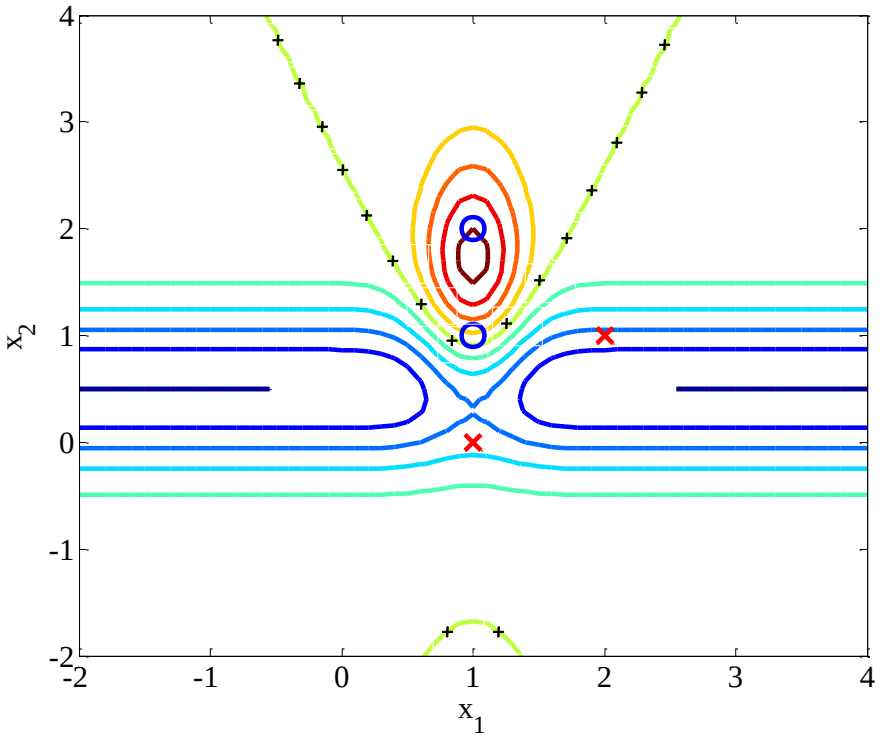


Figure 5.2 The contour plot of the difference surface $z = z_1 - z_2$

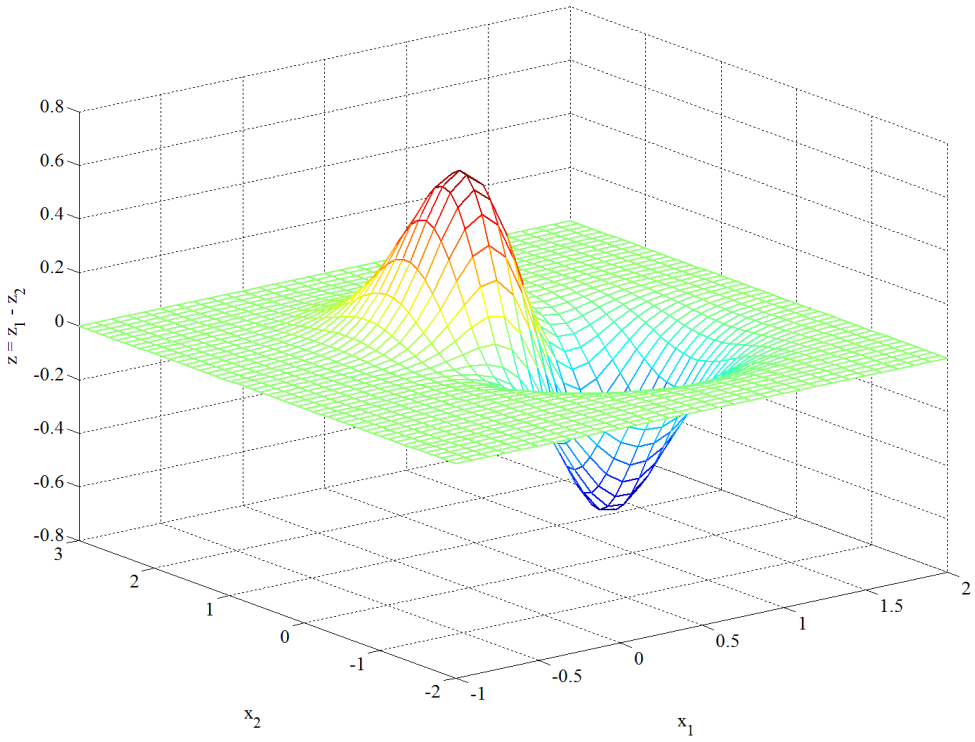


Figure 5.3 The mesh plot of the difference surface $z = z_1 - z_2$

Example 5.2 We model the same classes given in the Example 5.1 in the range space of $\mathbf{\Phi}_T$ using the method in Section 5.2.

The mesh plot and the contour plot of the difference surface $z = z_1 - z_2$ in the normalized space are shown in Figure 5.3 and Figure 5.4 respectively. The curve represented by "$- + -$" shown in Figure 5.4 is the decision boundary.

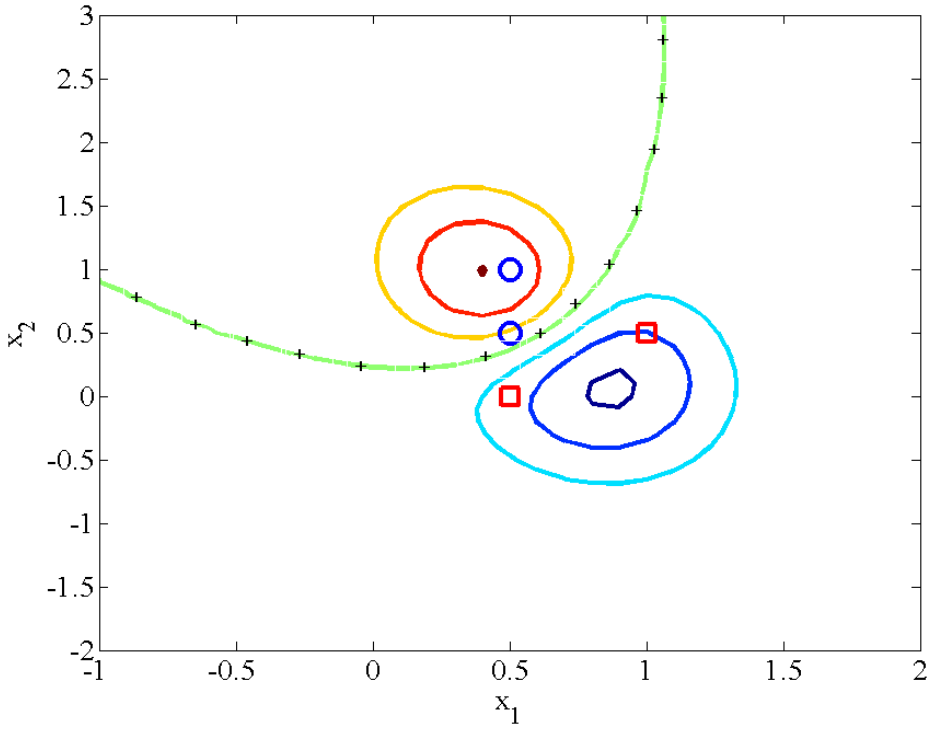


Figure 5.4 The contour plot of the difference surface $z = z_1 - z_2$

5.4 Experimental Work

In the experimental work, we compare the performances of the methods defined in Section 5.1, Section 5.2 and Support Vector Machines [81] in YALE, ORL, and AR face databases. The images automatically cropped according to the eye coordinates, resized to 66×60, 40×32, and 50×37 respectively. Support Vector Machine finds a hyperplane which optimally separates the *n*-dimensional

space into two categories. If the data cannot be separated by a linear hyperplane, SVM uses kernel functions to map the data onto a higher dimensional space to make it linearly separable. In the experiments, we used linear and quadratic kernels.

We perform pairwise classification [82] in the experiments. Pairwise classification converts the multi-class problems into series of two-class classification problems [83,84]. n-class classification problem is converted into $\frac{n(n-1)}{2}$ two-class classification problems. The experiments are executed the range space of the total within-class scatter matrix $\boldsymbol{\Phi}_T$ given in (2.18). The dimension of $\boldsymbol{\Phi}_T$ is equal to $M(N-1)$. The images are projected onto the range space of $\boldsymbol{\Phi}_T$ then normalized. In each experiment, we randomly selected N samples (N =5 for YALE, ORL and $N = 7$ for AR database) from each of M classes and the remaining samples are used for testing proposes. This procedure is repeated 5 times and the recognition rates are obtained by averaging each run. The experiments are executed in YALE, ORL, and AR face databases. The experimental results and their standard deviations are shown in Table 5.1 – Table 5.6. Here SVM-Lin denotes Support Vector Machines with linear kernel and SVM-Quad denotes Support Vector Machines with quadratic kernel. In the training stage SVM-Quad always gives 100% recognition rate. Similarly the method which is given in Section 5.2 gives close results to SVM-Quad. The best recognition results are achieved with the method given in Section 5.2 except one experiment. These results show that the proposed method given in Section 5.2 is successful against all of the other three methods.

Table 5.1 The recognition performance of the methods and their standard deviations in the training set on the YALE face database.

M	Method in Sec.5.1	Method in Sec.5.2	SVM-Lin	SVM-Quad
5	94.4 ± 4.6	**100 ± 0**	93.6 ± 4.6	**100 ± 0**
10	97.2 ± 2.7	98.8 ± 1.1	98.8 ± 1.1	**100 ± 0**
15	97.1 ± 1.7	98.7 ± 1.3	98.6 ± 1.3	**100 ± 0**

Table 5.2 The recognition performance of the methods and their standard deviations in the training set on the ORL face database.

M	Method in Sec.5.1	Method in Sec.5.2	SVM-Lin	SVM-Quad
10	99.6 ± 0.9	99.6 ± 0.9	72.0 ± 9.1	**100** ± 0
15	**100** ± 0	99.5 ± 0.7	72.0 ± 4.7	**100** ± 0
20	**100** ± 0	99.6 ± 0.5	69.0 ± 5.0	**100** ± 0

Table 5.3 The recognition performance of the methods and their standard deviations in the training set on the AR face database.

M	Method in Sec.5.1	Method in Sec.5.2	SVM-Lin	SVM-Quad
10	98.3 ± 1.9	**100** ± 0	95.0 ± 2.5	**100** ± 0
15	97.7 ± 0.9	**100** ± 0	92.6 ± 4.1	**100** ± 0
20	98.3 ± 0.4	**100** ± 0	92.7 ± 3.8	**100** ± 0

Table 5.4 The recognition performance of the methods and their standard deviations in the test set on YALE face database.

M	Method in Sec.5.1	Method in Sec.5.2	SVM-Lin	SVM-Quad
5	72.8 ± 16.8	78.4 ± 4.6	**81.6** ± 6.1	76.8 ± 6.6
10	58.4 ± 7.8	**72.0** ± 1.4	69.2 ± 5.0	63.6 ± 4.3
15	55.2 ± 3.1	**69.9** ± 2.0	65.3 ± 4.2	61.1 ± 1.7

Table 5.5 The recognition performance of the methods and their standard deviations in the test set on ORL face database.

M	Method in Sec.5.1	Method in Sec.5.2	SVM-Lin	SVM-Quad
10	70.8 ± 4.6	**86.4** ± 3.3	66.8 ± 12.5	80.0 ± 5.1
15	76.3 ± 2.4	**89.1** ± 3.0	65.6 ± 7.5	82.4 ± 6.1
20	71.2 ±3.0	**85.8** ± 6.2	59.6 ± 5.4	77.4 ± 7.5

Table 5.6 The recognition performance of the methods and their standard deviations in the test set on AR face database.

M	Method in Sec.5.1	Method in Sec.5.2	SVM-Lin	SVM-Quad
10	60.9 ± 4.4	**78.0** ± 6.5	65.4 ± 4.9	75.1 ± 6.2
15	57.7 ± 2.9	**76.8** ± 3.8	54.1 ± 2.1	73.7 ± 2.2
20	52.5 ± 4.5	**73. 9** ± 1.9	46.9 ± 1.9	69.4 ± 2.4

5.5 Summary of Covariance Estimation

In this section we give two class modeling methods in the range space of total within-class scatter matrix $\boldsymbol{\Phi}_T$. At first we project all samples in the range space of $\boldsymbol{\Phi}_T$ to reduce the dimensions.

In the first method, we generate the exponential surfaces to all classes. We define a diagonal rotation matrix $\boldsymbol{R}_c$. The variables of $\boldsymbol{R}_c$ are found using the steepest descent method. Since we use the diagonal rotation matrices, the level curves of the class model functions are the hyper ellipsoids which are parallel to the coordinate axis. The main reason using diagonal matrix is the difficulty in computation of variables. This method is not feasible for the databases which have numerous samples.

In the second method, we form new covariance matrices in the range space of $\boldsymbol{\Phi}_T$. We normalized the projected samples before modeling the classes. The inverse of the new covariance matrices are used to model the classes using exponential surfaces.

The within-class covariance matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$in the range space of $\boldsymbol{\Phi}_T$ are not full rank. Then the classes cannot be bounded in some dimensions. When we model a class, we use data of the both classes. The modified covariance matrices given in (5.15) and (5.16) are full rank so they are invertible. Since the classical quadratic classifiers use the inverse of within-class covariance matrix, they cannot be used in the range space of $\boldsymbol{\Phi}_T$.

We perform the experiments on the YALE, ORL and AR face databases with different number of classes. In the experiments we used pairwise

classification method. Experimental results show that the method described in Section 5.2 always gives better results except one case than the method given in Section 5.1, SVM-Lin and SVM-Quad.

6 CONCLUSION

In this thesis three problems in pattern recognition are examined namely, feature selection, single image per class problem, and within-class scatter matrix estimation in high dimensional space.

A novel feature selection algorithm is proposed that use the projection matrix of the common vectors. The importance of the features is determined by the corresponding column norms of the projection matrix of the common vectors of all classes. The most important parts of the face image for the recognition purposes are eyes, mouth, and nose. In the experiments it is seen that this assumption is correct. The experiments are performed on AR, ORL, and YALE face databases. We achieved great dimensionality reduction with small decrease in recognition rates. In the experiments with occluded face images we achieved not only great dimensionality reduction but also an increase in recognition rates. Also in the experiments carried out in YALE face database, we achieved great dimensionality reduction with slight increase in recognition rates. One of the most important conclusions of this work is the selection of the training set images. We think that the training set images must have variable backgrounds to eliminate the pixels belonging to the background. In ORL face database experiments, it is seen that our pixel elimination method is sensitive to face rotations. Also we proved that the importance of the pixels is independent of the selection of the basis vectors of the range space of the covariance matrix.

Collecting samples for a subject is a difficult task for face recognition applications. Well-known face recognition techniques fail if only one sample available for a subject. Many algorithms are proposed to overcome this difficulty. A novel image decomposition method using QRCP algorithm is proposed in our study. It is known that DCVA is an extension of FLDA. Also a two dimensional extension of DCVA is proposed. The performances of 2D-FLDA, DCVA, and 2D-DCVA are compared in one sample problem in five different face databases. Proposed decomposition method gave satisfactory results compared with SVD based decomposition algorithm. Also 2D-DCVA gave superior results than DCVA

and 2D-FLDA. Also it is seen that QRCP-based image decomposition method generally gives better results than SVD-based image decomposition method.

Covariance matrix estimation is an important problem especially in high dimensional space because of insufficient number of data. We proposed two within-class covariance matrix estimations. In both of the methods the data of classes are projected onto the range space of the total within-class scatter matrix, $R(\mathbf{\Phi}_T)$, which makes a great dimensionality reduction. In $R(\mathbf{\Phi}_T)$, the within-class covariance matrix of a class is modeled using not only its own data but also the data of the other classes. Experimental results show that the method described in Section 5.2 gives better results than SVM in $R(\mathbf{\Phi}_T)$.

BIBLIOGRAPHY

[1] Zhao, W., Chellappa, R., Phillips, P.J. and Rosenfeld, A., "Face Recognition: A Literature Survey," *ACM Computing Surveys*, **35** (4), 399-458, 2003.

[2] Abate, A.F., Nappi, M., Riccio, D. and Sabatino, G., "2D and 3D Face Recognition: A Survey," *Pattern Recognition Letters*, **28** (14), 1885-1906, 2007.

[3] Guyon, I. and Elisseeff, A., "An introduction to variable and feature selection," *Journal of Machine Learning Research*, **3**, 1157-1182, 2003.

[4] Theodoridis, S. and Koutroumbas, K., *Pattern Recognition*, Academic Press, USA, 1999.

[5] Jain, A.K., Duin, R.P.W. and Mao, J., "Statistican pattern recognition: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22** (1), 4-37, 2000.

[6] Dash, M. and Liu, H., "Feature Selection for Classification," *Intelligent Data Analysis*, **1**, 131-156, 1997.

[7] Jain, A. and Zongker, D., "Feature Selection: Evaluation, Application, and Small Sample Performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19** (2), 153-158, 1997.

[8] Saeys, Y., Abeel, T. and Peer, Y.V.d., "Robust feature selection using ensemble feature selection techniques," *Machine Learning and Knowledge Discovery in Databases*, **5212**, 313-325, 2008.

[9] Saeys, Y., Inza, I and Larranaga, P., "A review of feature selection techniques in bioinformatics," *Bioinformatics*, **23** (19), 2507-2517, 2007.

[10] Koç, M. and Barkana, A., "Feature selection with discriminative common vector approach in face recognition," *Pattern Recognition Letters*, (Under review).

[11] Raudys, S.J. and Jain, A.K., "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13** (3), 252-264, 1991.

[12] Liu, H. and Motoda, H., *Computational Methods of Feature Selection*, Chapman & Hall/Crc Data Mining and Knowledge Discovery Series, 2008.

[13] Tan, X., Chen, S., Zhou, Z.-H. and Zhang, F., "Face recognition from a single image per person: a survey," *Pattern Recognition*, **39** (9), 1725-1745, 2006.

[14] Gao, Q.-x., Zhang, L. and Zhang, D., "Face recognition using FLDA with single training image per person," *Applied Mathematics and Computation*, **205** (2), 726-734, 2008.

[15] Koç, M. and Barkana, A., "A new solution to one sample problem in face recognition using FLDA," *Applied Mathematics and Computation*, **217** (24), 10368-10376, 2011.

[16] Yin, H., Fu, P. and Meng, S., "Sampled FLDA for face recognitionwith single training image per person," *Neurocomputing*, **69** (16-18), 2443-2445, 2006.

[17] Zhang, D., Chen, S. and Zhou, Z.-H., "A New Face Recognition Method Based on SVD Perturbation for Single Example Image per Person," *Applied Mathematics and Computation*, **163**, 895-907, 2005.

[18] Wang, J., Plataniotis, K.N., Lu, J. and Venetsanopoulos, A.N., "On Solving the Face Recognition Problem with One Training Sample per Subject," *Pattern Recognition*, **39** (9), 1746-1762, 2006.

[19] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, 1991.

[20] Çevikalp, H., Neamtu, M., Wilkes, M. and Barkana, A., "Discriminative Common Vectors for Face Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **27** (1), 4-13, 2005.

[21] Belhumeur, P., Hespanha, J. and Kriegman, D., "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19** (7), 711-720, 1997.

[22] Chen, L.F., Liao, H.Y.M., Ko, M.T., Lin, J.C. and Yu, G.J., "A New LDA-Based Face Recognition System Which Can Solve the Small Sample Size Problem," *Pattern Recognition*, **33** (10), 1713-1726, 2000.

[23] Swets, D.L. and Weng, J., "Using Discriminant Eigenfeatures for Image

Retrieval," *IEEE Transcations on Pattern Analysis and Machine Intelligence*, **18** (8), 831-836, 1996.

[24] Liu, W., Wang, Y., Li, S.Z. and Tan, T., "Null space approach of fisher discriminant analysis for face recognition," *8th European Conference on Computer Vision (ECCV) Workshop BioAW*, Prague, 2004, 32-44.

[25] Das, K. and Nenadic, Z., "An efficient discriminant-based solution for small sample size problem," *Pattern Recognition*, **42** (5), 857-866, 2009.

[26] Çevikalp, H., Barkana, B. and Barkana, A., "A comparison of the common vector and the discriminative common vector methods for face recognition," *9th World Multiconference on Systemics, Cybernetics, and Informatics (WMSCI)*, 2005.

[27] Kong, H., Teoh, E.K., Wang, J.G. and Venkateswarlu, R., "Two dimensional Fisher discriminant analysis: Forget about small sample problem," *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, 2005, 761-764.

[28] Koç, M. and Barkana, A., "An implementation of discriminative common vector approach Using matrices," *The Seventh International Multi-Conference on Computing in the Global Information Technology(ICCGI2012)*, Venice, 2012.

[29] Nhat, V.D.M. and Lee, S., "Discriminative common images for face recognition," *In proceedings of ICANN - Part I*, **3696**, 563-568, 2005.

[30] Gulmezoglu, M.B., Keskin, M., Dzhafarov, V. and Barkana, A., "A Novel Approach to Isolated Word Recognition," *IEEE Trans. on Speech and Audio Processing*, **7** (6), 620-628, 1999.

[31] Gulmezoglu, M.B., Dzhafarov, V. and Barkana, A., "The Common Vector approach and its Relation to Principal Component Analysis," *IEEE Trans. Speech and Audio Processing*, **9** (6), 655-662, 2001.

[32] Koç, M., Barkana, A. and Gerek, Ö.N., "A Fast Method for the Implementation of Common Vector Approach," *Information Sciences*, **180** (20), 4084-4098, 2010.

[33] Günal, S. and Edizkan, Rifat, "Subspace Based Feature Selection for Pattern Recognition," *Information Sciences*, **178** (19), 3716-3726, 2008.

[34] He, Y., Zhao, L. and Zou, C., "Face recognition using common faces method," *Pattern Recognition*, **39** (11), 2218-2222, 2006.

[35] Edizkan, R., Gülmezoğle, M.B., Ergin, S. and Barkana, A., "Improvements on common vector approach for multi class problems," *13th European Signal Processing Conference*, 2005.

[36] Tamura, A. and Zhao, Q., "Rough common vector: a new approach to face recognition," *IEEE International Conference on Systems, Man, and Cybernetics*, 2007, 2366-2371.

[37] Gulmezoglu, M.B., Dzhafarov, V., Edizkan, R. and Barkana, A., "The common vector approach and its comparison with other subspace methods in case of sufficient data," *Computer Speech and Language*, **21** (2), 266-281, 2007.

[38] Çevikalp, H., Neamtu, M. and Barkana, A., "The kernel common vector method: a novel nonlinear subspace classifier for pattern recognition," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, **37** (4), 937-951, 2007.

[39] Naseem, I., Togneri, R. and Bennamoun, M., "Linear REgression for Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32 (**11), 2106-2112, 2010.

[40] Diaz-Chito, K., Ferri, F.J. and Diaz-Villanueva, W., "Image Recognition through Incremental Discriminative Common Vectors," *Anveced Concepts for Intelligent Vision Systems, LNCS*, **6475**, 304-311, 2010.

[41] Travieso, C.M., Botella, P., Alonso, J.B. and Ferrer, M.A., "Discriminative Common Vector for Face Identification," *43rd Annual Conference on Security Technology*, 2009, 134-138.

[42] Çevikalp, H., Neamtu, M. and Wilkes, M., "Discriminative Common Vector Method With Kernels," *IEEE Transactions on Neural Networks*, **17** (6), 1550-1565, 2006.

[43] Lakshmi, C. and Ponnavaikko, M., "Boosting Kernel Discriminative Common Vectors for Face Recognition," *Journal of Computer Science*, **5** (11), 801-810, 2009.

[44] Diaz-Chito, K., Ferri, F.J. and Diaz-Villanueva, W., "An Emprical Evaluation of Common Vector Based Classification Methods and Some Extensions," *Structural, Syntatic, and Statistical Pattern Recognition, LNCS*, **5342**, 977-985, 2008.

[45] Turk, M. and Pentland, A., "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, **3** (1), 71-86, 1991.

[46] Duda, R.O., Hart, P.E. and Strok, D.G., *Pattern Classification*, 2nd ed., Wiley-Interscience, 2001.

[47] Matrinez, A.M. and Kak, A.C., "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23** (2), 228-233, 2001.

[48] Koç, M. and Barkana, A., "Feature selection method with common vector and discriminative common vector approaches," *19th Conference on Signal Processing and Communications Applications (SIU)*, Antalya, 2011, 98-101.

[49] Choi, S.-I., Oh, J., Choi, C.-H. and Kim, C., "Input variable selection for feature extraction in classification problems," *Signal Processing*, **92** (3), 636-648, 2012.

[50] Koç, M. and Barkana, A., "Ayırtedici ortak vektör yaklaşımının tek örnek problemine uygulanması," *20. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SiU2012)*, Muğla, 2012.

[51] Brunelli, R. and Poggio, T., "Face Recognition: Features versus Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15** (10), 1042-1052, 1993.

[52] Chen, S., Liu, J. and Zhou, Z.-H., "Making FLDA applicable to face recognition with one sample per person," *Pattern Recognition*, **37** (7), 1553-1555, 2004.

[53] Xiong, H., Swamy, M.N.S. and Ahmad, M.O., "Two-dimensional FLD for face recognition," *Pattern Recognition*, **38** (7), 1121-1124, 2005.

[54] Su, K.-Y. and Lee, C.-H., "Speech recognitionusing wieghted HMM and subspace projection approaches," *IEEE Transactions on Speech and Audio Processing*, **2** (1), 69-79, 1994.

[55] Wang, X. and Paliwal, K.K., "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern Recognition*, **36** (10), 2429-2439, 2003.

[56] Yang, J., Frangi, A.F., Yang, J.-y., Zhang, D. and Jin, Z., "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27** (2), 230-244, 2005.

[57] Liu, C.-L., "High accuray Chinese character recognitino using quadratic classifiers with discriminative feature extraction," *18th International Conference on Pattern Recognition*, 2006, 942-945.

[58] Deepu, V., Madhvanath, S. and Ramakrishnan, A.G., "Principal component analysis for online handwritten character recognition," *17th International Conference on Pattern Recognition*, 2004, 327-330.

[59] Günal, S., Ergin, S., Gülmezoğlu, M.B. and Gerek, Ö.N., "On feature extraction for spam e-mail detection," *MCRCS2006, Lecture Notes in Computer Science*, **4105**, 635-642, 2006.

[60] Zheng, W., "Heteroscedastic feature extraction for texture classification," *IEEE Signal Processing Letters*, **16** (9), 766-769, 2009.

[61] Edwards, C.H. and Penney, D.E., *Elementary Linear Algebra*, Prentice Hall International, 1988.

[62] Seber, G.A.F., *Linear Regression Analysis*, Wiley-Interscience, 2003.

[63] Ryan, T.P., *Modern Regression Methods*, Wiley-Interscience, 1997.

[64] Kshirsagar, A.M., *Multivariate Analysis*, Marcel Dekker Inc., 1972.

[65] Li, M. and Yuan, B., "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, **26** (5), 527-532, 2005.

[66] Zheng, W.-S., Lai, J.H. and Li, S.Z., "1D-LDA vs. 2D-LDA: When is vector-based linear discriminant analysis better than matrix-based?," *Pattern*

Recognition, **41** (7), 2156-2172, 2008.

[67] Kong, H., Teoh, E.K., Wang, J.-G. and Kambhamettu, C., "Generalized 2D fisher discriminant analysis," *In Proceedings of the 16th British machine vision conference*, 2005, 1-10.

[68] Martinz, A and Benavente, R., *The AR Face Database*. CVC Technical Report, No: 24, 1994.

[69] *ORL Face Database*. AT&T Laboratories Cambridge, 1992-1994.

[70] Kanjilal, P.P., *Adaptive Prediction and Predictive Control*, Peter Peregrinus Ltd., 1995.

[71] Ari, S. and Saha, G., "In Search of an SVD and QRcp Based Optimization Technique of ANN for Automatic Classification of Abnormal Heart Sounds," *International Journal of Biological and Life Sciences*, **2** (1), 1-9, 2007.

[72] Chakroborty, S. and Saha, G., "Feature Selection Using Singular Value Decomposition and QR Factorization with Column Pivoting for Text-Independent Speaker Identification," *Speech Communication*, **52** (9), 693-709, 2010.

[73] Kailath, T., Sayed, A.H. and Hassibi, B., *Linear Estimation*, Prentice Hall, 1999.

[74] Meyer, C.D., *Matrix analysis and applied linear algebra*, SIAM, 2001.

[75] Yang, J., Zhang, D., Frangi, A.F. and Yang, J., "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," *IEEE Trannsactions on Pattern Analysisi and Machine Intelligence*, **26** (1), 131-137, 2004.

[76] Chowdhury, S., Sing, J., Basu, D. and Nasipuri, M., "Generalized Two-Dimensional FLD method for feature extraction: an application to face recognition," *Advances in Knowledge Discovery and Data Mining / Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2010, ch. 6119, 101-112.

[77] Phillips, P.J., Moon, H., Rizvi, S.A. and Rauss, P., "The FERET Evaluation Methodology for Face Recognition Algorithms," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, **22** (10), 1090-1104, 2000.

[78] Graham, D.B. and Allinson, G.N.M., "Characterizing Virtual Eigensignatures for General Purpose Face Recognition," *Face Recognition: From Theory to Applications*, (Ed: Wechsler, H. and ark.), NATO ASI Series / Computer and Systems Sciences, 1998, ch. 163, 446-456.

[79] Zhang, B., Zhang, L., Zhang, D. and Shen, L., "Directional Binary Code With Application to PolyU Near-Infrared Face Database," *Pattern Recognition Letters*, **31** (14), 2337-2344, 2010.

[80] Bishop, C.M., *Neural nerworks for pattern recognition*, Oxford University Press Inc., 1995.

[81] Schölkopf, B. and Smola, A.J., *Learning with Kernels*, MIT Press, 2002.

[82] Kreßel, U.H.-G, "Pairwise classification and support vector machines," *Advances in kernel methods: Support vector learning*, (Ed: Schölkopf, B., Burges, C.J.C. and Smola, A.J.), Cambridge, The MIT Press, 1999, 255-268.

[83] Krzyśko, M. and Wołyński, W., "New variants of pairwise classification," *European Journal of Operational Research*, **199** (2), 512-519, 2009.

[84] Park, S.-H. and Fürnkranz, J., "Efficient pairwise classification," *Proc. of the 17th European Conference on Machine Learning (ECML-07)*, 2007, 658--665.