Sparse Regression Based Face Recognition

Ahmad Jum'a M. Qudaimat

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical and Electronic Engineering

Eastern Mediterranean University
February 2019
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Assoc. Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Doctor of Philosophy in Electrical and Electronic Engineering.

Prof. Dr. Hasan Demirel
Chair, Department of Electrical and Electronic Engineering

We certify that we have read this thesis and that in our opinion, it is fully adequate, in scope and quality, as a thesis of the degree of Doctor of Philosophy in Electrical and Electronic Engineering

Prof. Dr. Hasan Demirel
Supervisor

Examining Committee

1. Prof. Dr. Tayfun Akgül
2. Prof. Dr. Hasan Amca
3. Prof. Dr. Hasan Demirel
4. Prof. Dr. Hüseyin Özkaramanlı
5. Prof. Dr. İlkay Ulusoy

ABSTRACT

Despite the large volume of research in the literature, face recognition remains a hard problem to solve. The challenge is due to many factors that may affect the performance of any recognition system such as size of training data, noisy images, accuracy-speed trade-off, variations in illumination, expressions, or pose, etc. Although there are a lot of efforts that have been proposed for special conditions, no method exists to work under unconstrained conditions with satisfactory performance.

In computer vision, human face classification is a very popular and important topic. This popularity comes from the wide-spread applications of face recognition such as entertainment, security, and control. Most, if not all, of these applications require the recognition systems to have low computational complexity and high recognition accuracy.

In this thesis, three methods are proposed for feature extraction and face classification. These methods are based on ℓ_2-norm regularized regression. The main idea of these methods is to train a dictionary capable of transforming a face image into a form that can be used to classify it into its correct class. Image representation in the transformation domain is a sparse vector with small number of nonzero coefficients. The goal is to come up with a transformation matrix such that the sparsity pattern depends on the class of the transformed image. This is accomplished by regression in addition to regularization terms and constraints. The three proposed ideas differ in how to attain this goal. The first proposed method uses the idea of predefined sparse matrix to specify the sparsity pattern of image transformation. The

second method constrains the transformation such that the inner product of the image transformation is minimized if the images are of different classes, and maximized if the images are of the same class. The last method transforms the images such that the nonzero coefficients do not overlap for different class, and the transformation of class images become close to the transformation of its mean vector.

Several simulation experiments are implemented and executed to test the performance and competence of the proposed methods. The simulations are performed with different benchmark face databases. The results prove that our proposed methods are distinguished over state-of-the-art methods by their accuracy, computational cost and robustness to image occlusion and corruption.

Keywords: Face recognition, sparsifying transform, dictionary learning, transform Learning, feature Extraction, regression.

ÖZ

Önerilen çok sayıda araştırmaya rağmen, yüz tanımanın çözülmesi zor bir problem olduğu görülmektedir. Buradaki zorluk, eğitim verilerinin boyutu, gürültülü görüntüler, doğruluk-hız karşilaştırılması, aydınlatmadaki değişimler, ifadeler veya pozlar gibi herhangi bir tanıma sisteminin performansını etkileyen birçok faktörden kaynaklanmaktadır. Özel şartlar için teklif edilen kısıtlayıcı olmayan koşullar altında tatmin edici bir performansla çalışmak için hiçbir yöntem mevcut değildir.

Bilgisayarlı görü alanında insan yüzü sınıflandırması çok popüler ve önemli bir konudur. Bu popülerlik eğlence, güvenlik ve kontrol gibi yaygın uygulamalardan gelmektedir. Bu uygulamaların tümü olmasa da çoğu tanıma sistemlerinin düşük hesaplama karmaşıklığına ve yüksek tanıma doğruluğuna sahip olmasını gerektirir.

Bu tezde, öznitelik çıkarımı ve yüz tanıma için üç yöntem önerilmiştir. Bu yöntemler l2-norm tabanlı düzenli regresyona dayanmaktadır. Ana fikir, bir yüz görüntüsünü doğru sınıfa sınıflandırmak için kullanılabilecek bir forma dönüştürebilecek bir sözlük oluşturmaktır. Dönüşüm alanındaki görüntü temsili, az sayıda sıfır olmayan katsayılı seyrek bir vektördür. Amaç, seyreklik modelinin dönüştürülen görüntünün sınıfına bağlı olduğu şekilde bir dönüşüm matrisi oluşturmaktır. Bu, düzenleme şartlarına ve kısıtlamalarına ek olarak, regresyon ile gerçekleştirilir. Önerilen üç fikir bu hedefe nasıl ulaşılacağı konusunda farklılıklar göstermektedir. Önerilen ilk yöntem, görüntü dönüşümünün seyrek kalıbını belirlemek için önceden tanımlanmış seyrek matris fikrini kullanır. İkinci yöntem, görüntüleri farklı sınıflarda ise görüntü dönüşümünün iç çarpımının en aza indirgenmesi ve görüntülerin aynı sınıfta olması

halinde en üst düzeye çıkarılması şeklinde dönüşümü sınırlar. Son yöntem, sıfır olmayan katsayıların farklı sınıflar için çakışmadığı ve sınıf imajlarının dönüşümü ortalama vektörünün dönüşümüne yakın hale geleceği şekilde görüntüleri dönüştürür.

Önerilen yöntemlerin performansını ve yeterliliğini test etmek için çeşitli simülasyon deneyleri uygulanmış ve sonuçlar üretilmiştir. Simülasyonlar, farklı yüz veritabanlarında gerçekleştirilmiştir. Sonuçlar, önerilen yöntemlerin son teknoloji yöntemlere göre doğruluk, hesaplama maliyeti, görüntü tıkanıklığı ve bozulmaya olan sağlamlığı ile ayırt edildiğini kanıtlamaktadır.

Anahtar Kelimeler: Yüz tanıma, seyrekleştirme dönüşümü, sözlük öğrenme, dönüşüm öğrenme, özellik çıkarma, regresyon.

DEDICATION

Dedicated to

My great parents

My beloved Wife

My lovely children Amro, Malak and Zeina

for their love, endless support and encouragement.

ACKNOWLEDGMENT

I would like to express my special appreciation and thanks to my supervisor Professor Dr. Hasan Demirel for his patience, motivation, and immense knowledge. I would like to thank him for encouraging my research and for allowing me to grow as a research scientist.

I would also like to thank my committee members, Prof. Dr. Hüseyin Özkaramanlı, Prof. Dr. Hasan Amca, Prof. Dr. İlkay Ulusoy, and Prof. Dr. Tayfun Akgül for their brilliant comments and suggestions.

My sincere thanks also goes to my friends, especially Dr. Hamza Ahmed who was always so helpful and provided me with his assistance throughout my dissertation.

A special thanks to my family. Words cannot express how grateful I am to my parents for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank my parents-in-law, my brothers and sisters for their constant support and encouragements. At the end I would like express appreciation to my beloved wife Sawsan who spent sleepless nights thinking in my progress and was always my support in every moment during my PhD research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\ell_2 - \text{norm}$	Euclidean norm
$\ell_1 - \text{norm}$	Manhattan norm
$\ell_0 - \text{norm}$	Number of non-zero elements
T	Transposition operator
$\mathbf{I}$	Identity matrix
μ	Average values
$E[.]$	Expected value
$\mathbf{C}$	Covariance matrix
W	Tranformation matrix
Y	Matrix representation of training images
X	Sparse Codes
$\|.\|_F$	Frobenius norm
$\|.\|_2$	Euclidean norm
$\|.\|_0$	$\ell_0 - \text{norm}$
$\circ$	Point-wise product
$\mathbf{1}$	Vector of ones
∇	Gradient

LIST OF ABBREVIATIONS

2D	Two Dimensional
2D DMWT	Two Dimensional Discrete Multiwavelet Transform
2D DWT	Two Dimensional Discrete Wavelet Transform
ANN	Artificial Neural Network
AR	Aleix Martinez and Robert Benavente
CRC	Collaborative Representation based Classification
CRP	Collaborative Representation based Projections
D	Dimensions
DCT	Discrete Cosine Transform
DL	Deep Learning
DWT	Discrete Wavelet Transformt
FDA	Fisher discriminant analysis
FSSL	Feature Selection and Subspace Learning
GWT	Gabor Wavelet Transform
ICA	Independent Component Analysis
ISVM	Improved Supported Vector Machine
LBP	Local Binary Patterns
LBPP	Local Binary Probabilistic Pattern
LDA	Linear Discriminant Analysis
LFW	Labeled Faces in the Wild
LPP	Locality Preserving Projection
MLP	Muti-Layer Perceptron

NMR	Norm-based Matrix Regression
NN	Neigherest neighbor
NP	Non-deterministic Polynomial-time
ORL	Olivetti Research Lab
PCA	Principal component analysis
SID	Stacked Image Descriptor
SPP	Sparsity Preserving Projections
SRC	Sparse representation based classifier
SRICE	Sparse Representation-based classification algorithm using Iterative Class Elimination
STLC	Sparsyfing transform for Image Classification
SVM	Supported Vector Machine
WSRC	weighted sparse representation
STLC	Sparsyfing transform for Image Classification
SVM	Supported Vector Machine
WSRC	weighted sparse representation

Chapter 1

INTRODUCTION

Face recognition is a field of computer vision that attempts to determine the individual's identity. The process works by capturing the image of a human face and compares it to images in a previously stored images database. This process is challenging due to many reasons such pose variations, lightening conditions, aging and other problems. Many methods are proposed in this field as will be explained in the subsequent chapters. In this chapter a brief introduction about the general face recognition problem is presented. The main objectives, challenges and contributions of the thesis are also explained. Finally the outline of the thesis is presented in the last section.

1.1 Objectives

The main objectives of any face recognition method is to develop an algorithm that can achieve the classification task reliably with high recognition rate under various conditions, runs fast with minimum computational complexity and requires low number of training images. Unfortunately, these objectives are still not being achieved even after decades of research. In this thesis, these objectives are decomposed into the following three smaller objectives

The first goal of the proposed methods is to improving the accuracy and the speed of the current face recognition algorithms. Even though the accuracy of face recognition

system is a vital requirement, it is not the only criterion that determines the preferred algorithm. In real applications, speed plays a crucial role. For example, in video surveillance system, it is required to process the face image and make a classification decision in very short time. So, the first objective of our thesis is to develop algorithms that are accurate and fast.

Our second objective is to be able to classify face images reliably under unconstrained conditions such as expression, illumination, changes of viewpoint, random occlusion and corruption. Despite the huge number of research over the past decades, this issue remains a challenge in face recognition. Therefore, one of the objectives of the proposed methods in this thesis is to achieve high recognition rates in the presence of these problems.

The third objective is to recognize images under the problem of lacking of training face images. Sometimes, it is costly to collect a large number of training images. In learning algorithms, low number of training data can increase the generalization error, which is known as over-fitting problem. So, the third objective is to achieve a satisfactory performance with low number of training images.

1.2 Motivations

Face recognition is a challenging task for computers and digital systems. In contrast, it is much easier process for human being. Recognizing people is an important and fundamental capability of our perception system. One can distinguish a person whom he knows among thousands of people. The ultimate goal is to understand he human perception system and to build a recognition system that approaches its easiness and accuracy.

In computer vision field, face recognition is one of the most important applications which have drawn the attention of researchers in the last decades for many reasons. One of the reasons is that it is widely used in commercial and security applications. Some examples of these applications are:

- Law enforcement: video surveillance, suspect tracking, post-event analysis, shoplifting.
- Smart cards: passports, national ID, drivers' licenses, voter registration.
- Information security: desktop logon, internet security.
- Entertainment: virtual reality, video games.

1.3 Problem Definition

Successful systems for face recognition need to address various problems. To the best of our knowledge, the following are the major challenges [2, 3] for face recognition systems:

- Pose variations: Face images are highly varied by changing the pose angles. The system accuracy dramatically degraded when these angles are large.
- Facial Expressions: Cause deformation of important facial features like eyebrows, eyes, nose, mouth and so reduce the recognition rates.
- Illumination variations: The images are vastly affected by surrounding illumination and cause a serious impairment on the performance of recognition systems.
- Aging: Since faces are greatly changing over the time, face recognition process becomes a challenging task even by humans.
- Large-scale system: Country's population could be hundreds of millions. A hard

question is how to build an efficient recognition system with huge database like that.

- Low resolution: Many cameras produce blur and low quality face images which degrade the recognition system efficiency.

- High dimensional space: Images are transformed into high dimensional vectors. The dimensions of these vectors are equal to the number of image pixels, which are usually a big number. The high dimensions increase the computational cost.

- Lack of training data: In order for an algorithm to work efficiently, it requires a lot of training images. Sometimes it is hard to have such number of training data, so it becomes a challenge to develop a system to work under this constraint.

Unfortunately, in realistic scenarios, these issues come in groups not singly. Grouped together make the face images of the same person severely varies, which results in a great degradation of the recognition system efficiency. Also, they make building an optimal system a hard task.

1.4 Contributions

In this thesis, three new ideas are proposed for face recognition.:

- An algorithm based on sparsifying transform is considered. It mainly learns a dictionary that can transform the image into sparse vectors. In the transformation domain, the images of the same class should have similar nonzero coefficient patterns that can be used for identification. The classification process of this method only requires transforming the image and makes norm comparisons. This thesis proposes a novel method in sparsity based image identification that uses analysis dictionaries unlike the conventional sparsity based methods. One advantage of the proposed algorithm

is the low computational cost of the classification process.

- A new ℓ_2-norm regularized regression based face image recognition method is proposed, with ℓ_0-norm constraint to ensure sparse projection. The proposed method aims to create a transformation matrix that transforms the images to sparse vectors with positions of nonzero coefficients depending on the image class. The classification of a new image is a simple process that only depends on calculating the norm of vectors to decide the class of the image.

- A new ℓ_2-norm based regression feature extraction and face recognition method is proposed. The method aims to train a transformation matrix that can transform images into sparse vectors in a way that can be used for classification. Sparseness of image representation is guaranteed by ℓ_0-norm constraint. While transforming the images is considered as a feature extraction process, the proposed method also presents a classification method based on the sparsity pattern of image transformation.

1.5 Thesis Outline

The next chapters of this thesis are organized in the following way: Chapter 2 presents a literature survey of methods related to the thesis topic. Transform domain methods and other known algorithms are reviewed. Most popular dimensionality reduction methods are also presented. Chapter 3 explains the methodology of the thesis where the image preprocessing steps are presented. Face image databases and challenges are also explained in this chapter. In chapter 4, the first proposed method for face image classification [4] is explained, computational complexity and experimental results also provided. Chapter 5 describes a sparse regression based method for face recognition. Experimental results are presented to test efficiency of the method under image occlusion and corruption. Chapter 6, also, provides a novel

regression based method with regularization for face image classification. This chapter demonstrates various experiments on benchmark database to show its efficiency in face image classification. Chapter 7 discusses conclusions and ideas for future work and research.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

In this chapter we will briefly review the main feature extraction and face recognition methods. Throughout the following sections we will consider the subspace methods, transform domain methods, neural network methods, sparsity based methods, and support vector machines.

2.2 Subspace Methods

Principal Component Analysis (PCA) [5] and Linear Discriminant Analysis (LDA) [6] are the most used techniques for face recognition. Both of these methods can be used for dimensionality reduction and for feature extraction and image recognition. More explanations on PCA and LDA are given on the next subsections.

Locality Preserving Projection (LPP) [7] method is an alternative approach for the well-known PCA method. LPP is a linear subspace method that seeks for the best linear approximation for eigenfunctions of Laplace operator on the face image manifold. LPP is used to reduce the dimensionality and for feature selection, it can also be used recognition [7, 8]. LPP was used to extract the local features in [9] where the effect of variations in illumination face expressions and pose are reduced. Many articles propose methods for face recognition with LPP [10–14]. In [15], Gu et al proposed the joint feature selection and subspace learning (FSSL) which was based on locality-

preserving projection.

Independent Component Analysis (ICA), another subspace approach for recognition, is an extension of PCA. Unlike PCA, it uses the high order statistics of the data [16–19]. ICA is a successful method for facial representation, feature extraction and image recognition [16, 20]. The basis images obtained by ICA should be independent while the basis images in PCA are uncorrelated. ICA introduced in [21, 22] for face recognition.

Many regression based methods that use $\ell_{2,1} -$ norm regularization were recently developed including Nuclear norm-based matrix regression for face classification (NMR) [23]. Nuclear norm based principal component analysis (N-2D-PCA) [24] and others [25–36].

2.2.1 Linear Discriminant Analysis (LDA)

LDA is a supervised classification method used often in pattern recognition and statistics. It looks for a linear combination that can separate different classes. Unlike PCA, LDA is designed for, in addition to dimensionality reduction, efficient classification. PCA is set mainly for perfect reconstruction.

LDA and PCA are closely related in the sense that both of them look for linear transformation that minimizes the mean-squared-error between the original input vectors and the transformed version of these vectors onto reduced dimensional space. The objective of LDA is to reduce variances of the signals or data points that belong to the same class and increase the variances between vectors that belong to different classes. More formally, LDA transform the vectors such that the between-class scattering is maximized and the within-class scattering is minimized [37].

Assume we have n input samples from K different classes, each class contains n_k samples. For these data samples, the method uses the scatter matrices $\mathbf{S_W}$ and $\mathbf{S_B}$, respectively. Where

$$\mathbf{S_W} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_i^k - \mu_k)(x_i^k - \mu_k)^T \tag{2.1}$$

and

$$\mathbf{S_B} = \sum_{k=1}^{K} n_k (\mu_k - \mu)(\mu_k - \mu)^T \tag{2.2}$$

where x_i^k denotes the i^{th} vector in class k. μ_k is class k mean vector and μ is the mean of all samples.

LDA objective function is

$$\begin{aligned} P_{opt} &= \underset{P}{\text{argmax}} \frac{P^T S_B P}{P^T S_W P} \\ &= [p_1, p_2, \cdots, p_m] \end{aligned} \tag{2.3}$$

The solution to this optimization problem is the generalized eigenvectors $p_i, i = 1, \cdots, m$ of $\mathbf{S_W}$ and $\mathbf{S_B}$ corresponding to the eigenvalues $\lambda_i, i = 1, \cdots, m$ as follows

$$\mathbf{S_B p_i} = \lambda_i \mathbf{S_W p_i}, \;\; i = 1, \cdots, m \tag{2.4}$$

To reduce the dimensionality only keep the first d eigenvectors corresponding to the maximum eigenvalues. After the projection of signals onto the space of these eigenvectors, a discriminating algorithm like SVM and NN can be used for efficient classification of the input query signals. Figure 2.1 shows a comparison between PCA and LDA. PCA the direction of the maximum variances while LDA considers the separability of the classes.

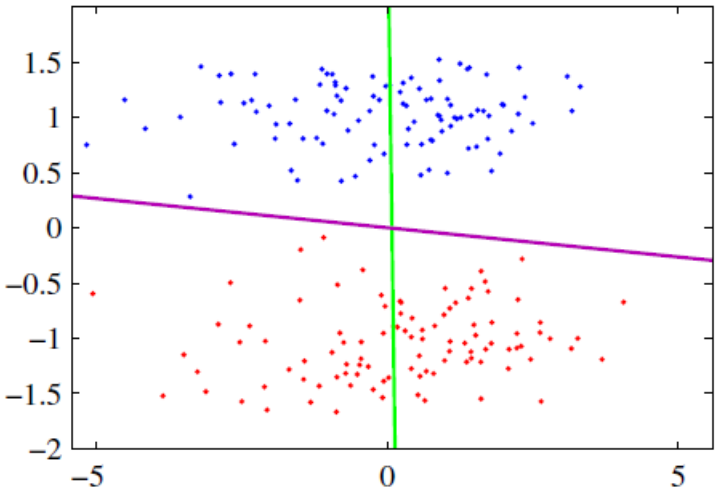


Figure 2.1: A comparison of PCA and LDA for dimensionality reduction [1]. Two dimensional data set from two different classes, shown in blue and red, to get projected onto one dimension. PCA selects the magenta curve which is the maximum variance direction. LDA considers the two class labels, so it selects the green curve.

2.3 Transform Domain Methods

Discrete wavelet transform (DWT) is successfully used for image recognition. Authors in [38] proposed wavelet filters for face recognition. DWT can integrated with other methods to achieve higher performance like PCA/DWT [39, 40], and ICA/2D DWT [41, 42]. DWT is extended to Discrete Multi-wavelet Transform (DMWT) which can be used for face recognition [43–47].

In spite of the fact that Discrete Cosine Transform (DCT) is an algorithm used extensively in data compression, it can also be used for face representation and recognition. DCT for face recognition was proposed in [48]. Face recognition by DCT integrated with Improved Supported Vector Machine (ISVM) was proposed in [49]. Many other researches proposed the integration of DCT with other methods such as DCT with GWT [50], DCT with NN [51], and DCT Local Binary Probabilistic Pattern (LBPP) [52].

2.4 Artificial Neural Network (ANN) Methods

Deep learning (DL) and ANN are used to increase the performance of face recognition. Multimodal facial representation based on DL [53] improves the accuracy of face identification system. Lei et al. [54] proposed the Stacked Image Descriptor (SID). Authors in [55] applied multilayer perceptron NN on the extracted features. The results is enhanced by applying DL NN instead multilayered perceptron [56]. DL and PCA are integrated to build an efficient face recognition algorithm [57].

2.5 Sparsity based Methods

Sparse Representation-based Classification (SRC) [58] can be considered as one of promising methods for face classification in the last few years. Detailed explanations of SRC are given in the next subsection. Some research articles are presented to analyze the mechanism of SRC [59, 60]. Qiao et al. [61] introduces a sparsity preserving projections (SPP) with SRC. Since sparsity based classification has been efficiently applied to face recognition, numerous researches proposed in this field like collaborative representation based classification (CRC) [60] and its collaborative representation based projections (CRP) [62] feature extraction method. Yang et al. [63] proposed a sparsity constrained regression for sparse coding. SRC-FDC [64] integrated Fischer criterion [65] with SRC. WSRC [66] proposed weights for the sparse coefficients proportional to the distance between the query image and the dictionary atoms. Sparse representation-based classification by iterative class-elimination (SRICE) [67] proposed an algorithm that iteratively eliminates number of classes corresponding to the maximum reconstruction error. Many other sparsity based methods for face recognition are proposed [68–78]

2.5.1 SRC Method

In SRC algorithm, face image is expressed as a combination of small number of atoms from face dictionary. This representation penalized by $\ell_1 - norm$ of the coefficient vector.

Given a set of training images $\mathbf{A}$ consist of face images form K classes. Let $\mathbf{A}_k = [\mathbf{A}_1, \mathbf{A}_2 \cdots, \mathbf{A}_K]$ where $\mathbf{A}^k = [\mathbf{v}_1^k, \mathbf{v}_2^k \cdots, \mathbf{v}_{n_k}^k] \in \mathbb{R}^{N \times n_k}$ is the matrix representation of class k images and $\mathbf{v}_i^k$ represent the i^{th} image that class. Class k query image $\mathbf{y} \in \mathbb{R}^N$ can be well represented by its class images in $\mathbf{A}_k$ as

$$\mathbf{y} = \sum_{i=1}^{n_k} a_i^k \mathbf{v}_i^k \tag{2.5}$$

using the matrix representation of all training images $\mathbf{A}$, equation 2.5 can be reformulated as

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 \tag{2.6}$$

where the entries of the coefficient vector $\mathbf{x}_0 = [0, \cdots, 0, a_1^k, a_2^k, \cdots, a_{n_k}^k, 0, \cdots, 0]^T$ are all zeros except those at positions associated with k^{th} class. In this case, if the coefficient vector corresponding to a test image is correctly given, the class of that image can be determined.

Although SRC uses the whole training images to represent the test image $\mathbf{y}$, it assumes that the image can be sufficiently represented by using only the training samples of the same class, which gives the sparsest possible solution among all. SRC looks for the sparsest possible solution for equation 2.6 by solving the optimization problem given below:

$$\hat{\mathbf{x}} = \underset{X}{\operatorname{argmin}} \|\mathbf{x}\|_0 \ \text{ subject to } \ \mathbf{A}\mathbf{x} = \mathbf{y} \tag{2.7}$$

If the number of nonzero coefficients in $\mathbf{x}$ is less than $N/2$, then the solution is unique [79]. Since $\ell_0 - minimization$is NP-hard, SRC uses the theories of compressed sensing and sparse representation in [80], [81] and [82] to solve this problem using $\ell_1 - minimization$ described as:

$$\hat{\mathbf{x}} = \underset{X}{\operatorname{argmin}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y} \tag{2.8}$$

After solving the optimization problem described above and finding the sparse vector $\hat{\mathbf{x}}$, the identification of the test image can be achieved by computing the residuals corresponding to each class, then selecting the one with the least value. computing the residual for class k can be computed as:

$$r_k(\mathbf{y}) = \|\mathbf{y} - \mathbf{A}\delta_k(\hat{\mathbf{x}})\|_2 \tag{2.9}$$

where $\delta_k : \mathbb{R}^N \to \mathbb{R}^{n_k}$ is a function that selects the coefficients $\hat{\mathbf{x}}$ that correspond to k^{th} class.

Solving equation 2.8 required that the dictionary $\mathbf{A}$ to be over-complete, in other words, the feature dimensions is much less than the number of training facial images, i.e. $N << n$. If this condition satisfied, the system becomes undetermined and has multiple solutions from which the sparsest one is selected. However, in reality, the dimensions of the facial images are very high and the number of training images is limited. So, one of the dimensionality reduction techniques like PCA [5], LDA [6], LPP [7] should be applied as

$$\tilde{\mathbf{y}} = \mathbf{R}\mathbf{y} = \mathbf{R}\mathbf{A}\mathbf{x} \in \mathbb{R}^d \tag{2.10}$$

where $\mathbf{R} \in \mathbb{R}^{d \times N}$ with $d << N$ is the dimensionality reduction matrix.

2.6 Support Vector Machines (SVM)

SVM [83] is a classification method that has been extensively studied and modified [84–88]. It consists of two phases; training phase and classification phase. In training phase, the method seeks a hyper-plane that separates the two classes and maximizes the marginal distance between the points from different classes. In classification stage, given a new data point, SVM finds out to which side of the separating hyper-plane it belongs and based on that SVM decides the class of the test point.

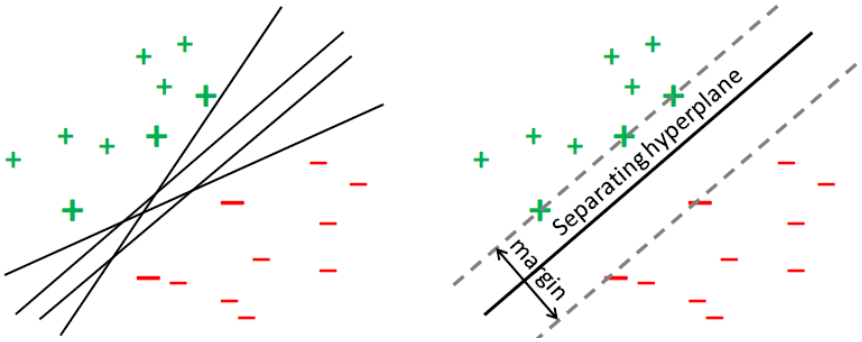


Figure 2.2: Illustration of SVM Classifier

Given n training points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, where $x_i \in \mathbb{R}^N$ represent the data point and $y_i \in \{-1, +1\}$. The separating hyper-plane can be defined as:

$$\{\mathbf{x} : f(\mathbf{x}) = \mathbf{x}^T \beta + \beta_0 = 0\} \tag{2.11}$$

where $\|\beta\| = 1$. SVM searches for the largest margin between these data points in class $+1$ and class -1. see figure 2.2. The optimization problem can be formulated as:

$$\begin{aligned} \min_{\beta, \beta_0} \quad & M \\ \text{subject to} \quad & y_i(\mathbf{x}_i^T \beta + \beta_0) \geqslant M, \;\; i = 1, \cdots, n \\ & \|\beta\| = 1 \end{aligned} \tag{2.12}$$

The derivation of this problem will not be considered here. Using the theories of

convex optimization, the solution to equation 2.12 for β can be expressed as:

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i \tag{2.13}$$

where the nonzero coefficients $\hat{\alpha}_i$ called support vector since $\hat{\beta}$ is represented in terms of them.

Given the vectors β and β_0, the classification function for this data set can be written as:

$$class(\mathbf{x}) = sign\left[\mathbf{x}^T\hat{\beta} + \hat{\beta}_0\right] \tag{2.14}$$

Having completed the literature review part of this thesis. We will present our own contributions in the next chapters.

Chapter 3

METHODOLOGY

In this chapter, section 3.1 presents a deeper look at the dimensionality reduction methods that are applied to the face images as a pre-processing step before implementing the face recognition methods. Description of the face databases and main challenges in each one are explained in section 3.2. Section 3.3 explains the performance evaluation criteria of the proposed methods.

3.1 Dimensionality Reduction

Real data signals, such as digital images, biomedical signals, and speech signals, have high dimensionality. This makes any processing to such kind of signals computationally expensive. In order to overcome this problem, its dimensionality needs to be reduced. Dimensionality reduction is mapping from high-dimensional representation of data to a meaningful lower dimensional space. It keeps the meaningful variation in the data and abandons the uninformative variances. Dimensionality reduction mitigates the undesired properties in the high dimensional spaces [89] and facilitates many operations on signal processing such as image classifications and compression.

Mathematically, if the original feature space is of dimension D contains n signals $x_1, x_2, \cdots, x_n$. Each $x_i \in \mathbb{R}^D$ is projected into a lower dimensional subspace to produce $y_1, y_2, \cdots, y_n$ where $y_i \in \mathbb{R}^d$. To reduce the dimensionality we need the new

dimensionality $d < D$. In matrix notation, let the matrix $\mathbf{X} \in \mathbb{R}^{D \times n}$ describes the input feature space such that the i^{th} column of $\mathbf{X}$ represent the signal x_i in the D-dimensional space. The linear projection from D to d dimensional space is given by $\mathbf{Y} = \mathbf{P}^T\mathbf{X} \in \mathbb{R}^{d \times n}$, where $\mathbf{Y}$ contains the n projected signals in $d-$dimensional space, and $\mathbf{P} \in \mathbb{R}^{D \times d}$ is the projection matrix.

3.1.1 Principal Component Analysis (PCA)

PCA is one of the most used methods for unsupervised dimensionality reduction [5, 90–92]. It makes use of the data redundancy in the training set, so it can be represented in a more compressed form. PCA projects the data vectors along the directions of maximum variances. Equivalently, it searches for the projections with minimum mean-squared distance between the vectors and their projections on the principal components.

Assume we have an n input data samples of dimension D, represented by the column of the input matrix $\mathbf{X} = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{D \times n}$. PCA transforms $\mathbf{X}$ into lower dimensional signals $\mathbf{Y}$ according to

$$\mathbf{Y} = \mathbf{P}^T(\mathbf{X} - \mu_{\mathbf{X}}) \tag{3.1}$$

where $\mu_{\mathbf{X}}$ is a vector of the average values of the input signals defined by the following relation

$$\mu_{\mathbf{X}} = E[\mathbf{X}] = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.2}$$

The projection matrix $\mathbf{P}$ in equation (3.1) is computed using the covariance matrix of the input signals $\mathbf{C}_{\mathbf{X}}$, of size $D \times D$. The covariance is a measure of the relative spread or variability, form the means, between two dimensions. Assuming $\mathbf{A}$ is the input matrix after subtracting the mean from each vector, i.e. $\mathbf{A} = [x_1 - \mu_{\mathbf{X}}, x_2 - \mu_{\mathbf{X}}, \cdots, x_n -$

$\mu_{\mathbf{X}}]$ It can be computed as

$$\begin{aligned} \mathbf{C_X} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathbf{x}})(x_i - \mu_{\mathbf{x}})^T \\ &= \mathbf{A}\mathbf{A}^T \end{aligned} \tag{3.3}$$

The eigenvectors of $\mathbf{C_X}$ form the basis vectors for the new subspace where we want to project the data points. The eigenvalues defines the variances after projection on its corresponding eigenvector. Since the size covariance matrix is $D \times D$, The number of its eigenvectors is D. The eigenvectors ,in $\mathbb{R}^D$, corresponding to the largest eigenvalues are called the principal components.

Columns of the matrix $\mathbf{P}$ are the principal components of the covariance matrix $\mathbf{C_X}$ in descending order based on the values of the corresponding eigenvalues. If we want the dimensionality of the new subspace to be d, then we will keep the first d eigenvectors.

When dealing with digital images, for example, the dimension D will be very high. and the computation of eigenvalues/eigenvectors of matrix $\mathbf{A}\mathbf{A}^T$ becomes computationally expensive and not possible in some cases. The alternate way to find them is by computing eigenvalues and eigenvectors of $\mathbf{A}^T\mathbf{A} \in \mathbb{R}^{n\times n}$. The nonzero eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are the same, they have at most $n-1$ nonzero eigenvalues. The eigenvectors of $\mathbf{A}\mathbf{A}^T$ can be found by the relation $\mathbf{V} = \mathbf{A}\mathbf{U}$ where $\mathbf{U}$ and $\mathbf{V}$ are the matrices of the eigenvectors of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$, respectively.

The representation coefficients $\alpha_{\mathbf{i}}$, in the transformation domain, for any input vector x_i can be computed as $\alpha_{\mathbf{i}} = \mathbf{P}^T(x_i - \mu_{\mathbf{X}})$. $\alpha_{\mathbf{i}}$ is a compact and unique representation of x_i. When the face images are used as input signals, it is referred to these eigenvectors

as eigenfaces [5].

3.2 Face Databases

Face databases are used to evaluate the performance of the face recognition algorithms. Since experiments on one database are not enough to show the performance of the recognition system, several databases should be used. The database should be big with large number of classes. The images should also be taken under various conditions, such as lightning conditions and illuminations, pose variations, facial expressions, and occlusion. Considering the previous remarks, we have selected four standard databases which are ORL [93], Extended Yale-B [94], AR [95], and LFW [96].

Figure 3.1: Sample images for one person in ORL database

3.2.1 ORL Database

ORL face database [93] consists of four hundreds facial-images from 40 males and females. ten face images are taken for each person with variances in illumination, expressions (open eyes / closed eyes, smiling / not smiling) and details (glasses/ no glasses). Each image is cropped to 92×112 pixels. Images are frontal view of the faces with rotation and tilting tolerance of 20°. Figure 3.1 shows the images of one person in ORL database.

3.2.2 Extended Yale-B Database

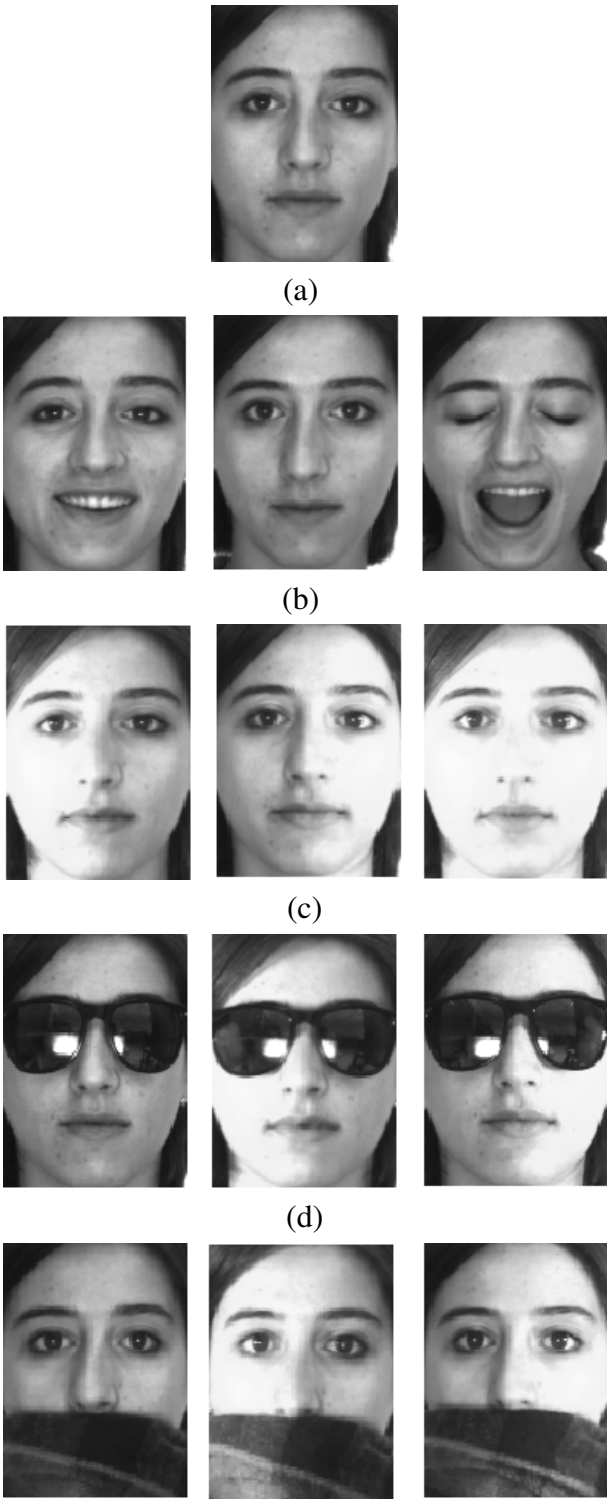
Extended Yale Face Database [97] consists of 2414 images for 38 subjects. Images are taken under 64 illumination conditions in frontal pose. They can be divided into 5 illumination groups according to the angle of light source. 60 images for each person are used. The images are cropped to 168×192 pixels. Figure 3.2 shows sample images for one person in this database.

Figure 3.2: Sample images for one person in ORL database

3.2.3 AR Database

AR database [95] has been created by Aleix Martinez and Robert Benavente. It has more than 4000 colored face images for 126 persons (70 men and 56 women), samples of AR images for one person are shown in figure 3.3. Frontal view images with various facial expressions as shown in figure 3.3(b), illumination conditions as shown in figure 3.3(c), and occlusion as shown in figure 3.3(d,e). The images are taken in two sessions which are separated by 14 days. No restrictions on make-up, hair style, or clothes.

From this database we randomly selects 100 persons for 50 men and 50 women. For each individual, 26 images, a subset of these images are selected for training and the remaining images are used for testing procedure. Each image is cropped to dimension 165×120 pixels.



(a)

(b)

(c)

(d)

(e)

Figure 3.3: Sample images for one person in AR database.

3.2.4 LFW Database

Labeled Faces in the Wild (LFW) is a face database that contains more than 13000 images collected from the web. Images are taken for 5749 persons. It has many

variations and large diversity.

In this thesis, we use the cropped version of this database LFWcrop, where only the center portion of the images are kept. The images are cropped to 64×64 pixels and the background is removed. From this database, we use 100 subjects. For each person, we select 6 images. Image samples for one person is shown in figure 3.4

Figure 3.4: Sample images for one person in LFW database

3.3 Performance Evaluation

In order to show the competitive performance of the proposed methods, They are tested and compared with the equivalent state-of-the art methods. The evaluation of the performance is measured by the recognition accuracy, where the recognition accuracy is the ratio between the number of the test images that are classified to their correct classes and the number of the test images that are incorrectly classified.

Stability and convergence of the optimization problem in each one of the proposed method is studied and. The stability and convergence are proved by performing simulations to show the progress of the value of the objective function with every iteration.

Chapter 4

SPARSIFYING TRANSFORM LEARNING FOR FACE IMAGE CLASSIFICATION

4.1 Introduction

This chapter presents a method for face identification based on sparsifying transform learning [98]. The main goal of the proposed method is to train a dictionary such that it can sparsify the image vectors in a pattern depending on its class. Based on this distinguishing pattern, the classifications process uses the transformation of the image to specify its class. The classification is simple and fast; it depends on comparing the norms of the parts of vector belonging to each class in the transform domain.

The next sections of this chapter are as follows. Details of the method are discussed in Section 4.2. Section 4.3 presents the solution procedure for the method. Classification process is explained in section 4.4 . Analysis of the computational cost of the method introduced in sec 4.5. Experimental validation with face image databases are discussed in Section 4.6. Summary of the conclusion in Section 4.7.

4.2 Proposed Method

In sparsity based approaches for image recognition, the test image is represented as a liner combination of trained synthesis dictionary. The atoms of the dictionary are arranged according to the classes they belong to. Although SRC achieves high recognition rates, $\ell_1 - minimization$ process is required each time a new image

needed to be identified, which impose a high computational cost.

To avoid the high cost of minimization problems with each identification process, Taking into account the dual relation between analysis and synthesis transformation, and we proposed sparsifying transform for image classification or STLC method, based on sparsifying transform proposed in [98]. We construct a dictionary capable of sparsifying the image in a way where the sparse coefficients are scattered in specific distribution depending on the image class. The algorithm uses this distribution pattern to classify the query images.

In classification process, the test image will be transformed by the trained dictionary and based on the distribution pattern and values of the nonzero coefficients in the transform domain, the method will decide to which class it belongs.

4.2.1 Problem Formulation and Objective Function

Given n training samples in $N-dimensional$ space, from K classes. k^{th} class represented in a matrix form as

$$Y^k = [y_1^k, y_2^k, \cdots, y_{n_k}^k] \in \mathbb{R}^{N \times n_k}, \;\; k = 1, 2, ..., K \tag{4.1}$$

here y_i^k and n_k are the i^{th} image and number of samples in class k, respectively. Here the superscript used to indicate class number. Denote Y as the matrix of all classes:

$$Y = [Y^1, Y^2, \cdots, Y^K] \in \mathbb{R}^{N \times n} \tag{4.2}$$

The objective is to create a sparsifying dictionary $W \in \mathbb{R}^{L \times N}$ appropriate to image classification. Besides the capability of W to sparsify any training image y_i, the

nonzero coefficients in the sparse transformed vector Wy_i should have a distinguishing pattern depending on the class of y_i.

Let the sparse code of WY be X. The question of creating the matrix W can be formulated as

$$(\text{P1}) \ \min_{W,X} \|WY - X\|_F^2, \ s.t. \ \|X_i\|_0 \leq s_i \ \forall i \tag{4.3}$$

where $\|X_i\|_0 \leq s$ constraint implies sparsity level of each columns of X. Since the purpose of the matrix W is image identification, we do not care about the dictionary capability of image reconstruction.

The sparse code X is computed by the element-wise product of the transformed images WY and the matrix $P \in \mathbb{R}^{L \times n}$.

$$X = WY \circ P \tag{4.4}$$

where $\circ$ denotes the element wise product. The matrix P determines the sparsity pattern in the transform domain. It has been introduced in this work in order to make the non-zero coefficients of the transformed training image to be in specific positions during the dictionary learning. Each column in P controls the transformation of one training image. Matrix P enforces the trained dictionary to transform the images that belong to one class to a specific pattern. There are various options for choosing the sparsity pattern matrix. In this work, the entries of P are zeros except non-overlapping sub-blocks of ones at its diagonal, which is shown in white in figure 4.1. Each sub-block $B_k \in \mathbb{R}^{s_k \times n_k}, \ k = 1, ..., K$ corresponds to one class, for example, the upper left white sub-block corresponds to the first class. As mentioned earlier, n_k is the number of class k training samples. s_k is the sparsity level of the vectors in X corresponding to class k.

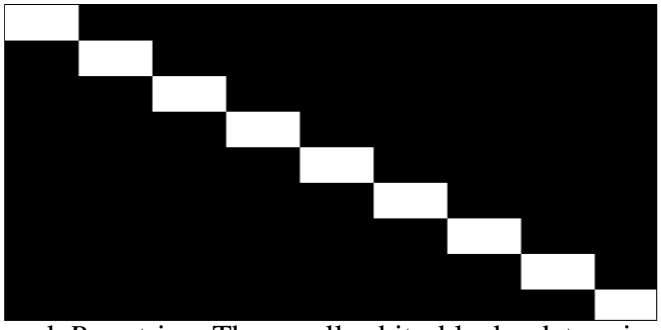


Figure 4.1: Cropped P matrix. The small white blocks determine the positions of sparse nonzero coefficients of the corresponding vector. Also the relative length of the white block gives indication of the sparsity level of each vector.

4.2.2 Trivial Solution

Although The optimization problem (P1) considers the fitting of the transform model to the data under the given sparsity constraint, it has a serious defect. It suffers from the trivial solution $W = 0$ and $X = 0$. Similar to the ideas in analysis dictionary learning [99, 100] and analysis dictionary learning [101, 102], we introduce additional penalties or constraints on the norm of W in the proposed minimization problem as follows.

$$\begin{aligned} (\text{P2}) \min_{W,X} \quad & \|WY - X\|_F^2 + \lambda_1 \|W\|_F^2 - \lambda_2 \log \|W\|_F^2 \\ \text{s.t.} \quad & \|X_i\|_0 \leq s_i, \ \forall i \end{aligned} \tag{4.5}$$

where λ_1 and λ_2 are the penalty parameters. The added penalty term $\lambda_1 \|W\|_F^2 - \lambda_2 \log \|W\|_F^2$ ensures that W will not go to zero. With these penalty functions, $\|W\|_F^2$ is bounded below by a value determined by the ratio λ_2/λ_1. For the purposes of image recognition, the actual value of the matrix norm is not an issue as long as the algorithm aims to do face recognition only.

4.3 Solution Procedure

The problem in equation (4.5) can be achieved iteratively by alternating between two steps; sparse coding and dictionary update.

4.3.1 Sparse Coding Step

Solve equation (4.5) by updating X while keeping W fixed

$$\min_{W,X} \|WY - X\|_F^2, \; s.t. \; \|X_i\|_0 \leq s, \; \forall i \tag{4.6}$$

the solution of the above equation for X can be found by point-wise product with the sparsity pattern matrix as in equation (4.4). This will preserve the coefficients in X that is in the same positions of ones in matrix P.

4.3.2 Dictionary Update Step

Solve equation 4.5 by updating W while keeping X fixed. Equation 4.5 is switched to following minimization problem

$$\min_{W} \|WY - X\|_F^2 + \lambda_1 \|W\|_F^2 - \lambda_2 \log \|W\|_F^2 \tag{4.7}$$

The optimization problem in equation (4.7) is non-convex. To illustrate this fact, figure 4.2 shows a plot of $\lambda_1 \|W\|_F^2 - \lambda_2 \log \|W\|_F^2$. W is a diagonal $2X2$ matrix. To clarify the figure, the parameters λ_1 and λ_2 were set to 1 and 20, respectively. Even though the figure proves the non-convexity of the optimization problem, it still can be solved using gradient descent algorithm with fixed or backtracking line search [103]. The gradient formulas for the terms in the objective function [104] are

$$\nabla_W \|W\|_F^2 = 2W \tag{4.8}$$

$$\nabla_W \|WY - X\|_F^2 = 2WYY^T - 2XY^T \tag{4.9}$$

$$\nabla_W \log \|W\|_F^2 = \frac{2W}{\|W\|_F^2} \tag{4.10}$$

The dictionary W can be initialized randomly and different stopping rules can be applied for the gradient algorithm iterations, such as the change in the norm of the

gradient, or the best recognition rate as in our case. However, the algorithm converges very quickly after few numbers of iterations as we will show in the experimental results section.

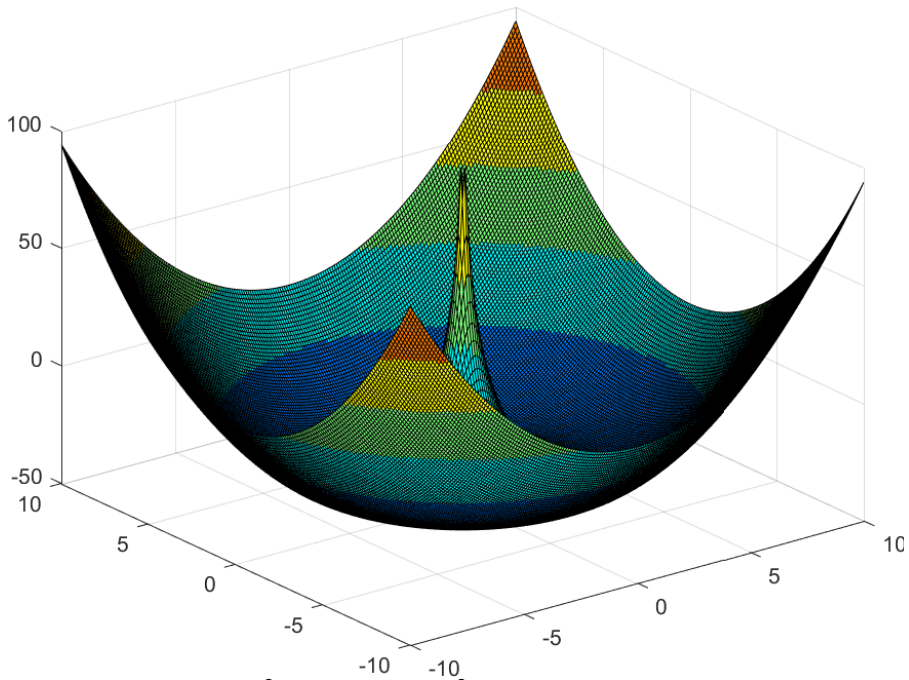


Figure 4.2: PLot of $\lambda_1 \|W\|_F^2 - \lambda_2 \log \|W\|_F^2$ of $2X2$ diagonal matrix W. The x and y axis indicate the first and second diagonal entries of W. $\lambda_1 = 1$, $\lambda_2 = 20$.

4.4 Classification

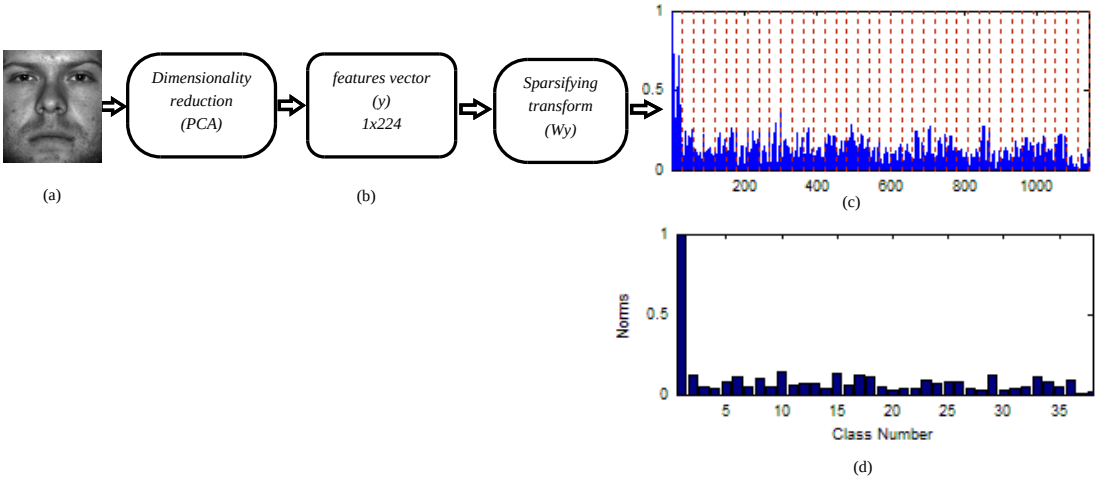
Given a new query image y that belongs to one class of the training subjects, the test image will first be transformed by the trained dictionary W as in the following equation

$$x = Wy \tag{4.11}$$

where the significant nonzero coefficients of the transformation vector x is expected to be concentrated in the positions that associated with the image class and small value coefficients in the positions that associated with the other classes as shown in Figure 4.3.

Assume that, for each class k, we have a function $\delta_k(x) : \mathbb{R}^L \to \mathbb{R}^{s_k}$, that selects only the coefficients in x corresponding to class k. Then the test image y will be assigned based on the minimum $\ell_2 - norm$ as

$$class(y) = \underset{k}{\operatorname{argmin}} \|\delta_k(x)\|_2^2 \tag{4.12}$$



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

Figure 4.3: Sparse coefficients. (a) Original image form the first class of extended Yale B face database of size 192×168. (b) Downsampled image to size 24×21. (c) The coefficients in the vector representation of the downsampled image in the transform domain, the coefficients are downsampled by factor of 4 for clear appearance. (d) The norms of coefficients belonging to each class.

4.5 Convergence and Computational Complexity

The algorithm described in (P2) can be achieved iteratively by alternating between two steps; sparse coding and dictionary update. The objective function is theoretically bounded below by λ_2/λ_1. The convergence in terms of recognition rate is proved empirically as we will see in the experimental result section.

We run the algorithm implemented by outer iterations for sparse coding. Inner iterations to calculate the gradient descent algorithm. YY^T computed only once during the whole process and requires N^2L product-sum operations.

The sparse coding step in each iteration involves the computation of $WY \circ P$ which requires approximately NL^2 product-sum operations. It also includes the computation of XY^T which requires αNL^2 operations, where α is the ratio of nonzero elements in matrix P. Thus the sparse coding iteration requires $(1+\alpha)NL^2$ operations.

In each dictionary update round, we need to compute WYY^T which requires $2NL^2$ product-sum operations and $\|W\|_F^2$ which requires NL operations. Thus the dictionary update step requires $NL(2L+1)$ operations.

Since the number of outer and inner iterations are fixed, then in total it requires $O(NL^2)$ product-sum operations. Where N, as mentioned earlier, is the space dimensionality, and L is the number of rows in W, which is less than or equal the number of training images.

4.6 Experimental Validation

We test our method with three benchmark databases; ORL [93], AR [95] and the extended-Yale [94] databases. 2-fold cross validation was used. The parameters λ_1, λ_2 were set to 1×10^6 and 1×10^8, respectively. The number of the outer and inner iterations were set to 200 and 100, respectively. The dictionary initialized as $W = Y^T$, so $L = n$. The simulations were carried out with Intel Xeon CPU at 2.2GHz and 8GB memory.

4.6.1 Stability and Convergence of The Proposed Method

Figure 4.4 shows the progress of the proposed algorithm over iterations. The value of main part of the objective function in optimization problem (P2), $\|WY - X\|_F^2$, is monotonically decreasing over iterations. At the first few iterations, the figure shows a fast change then it reaches its steady state value.

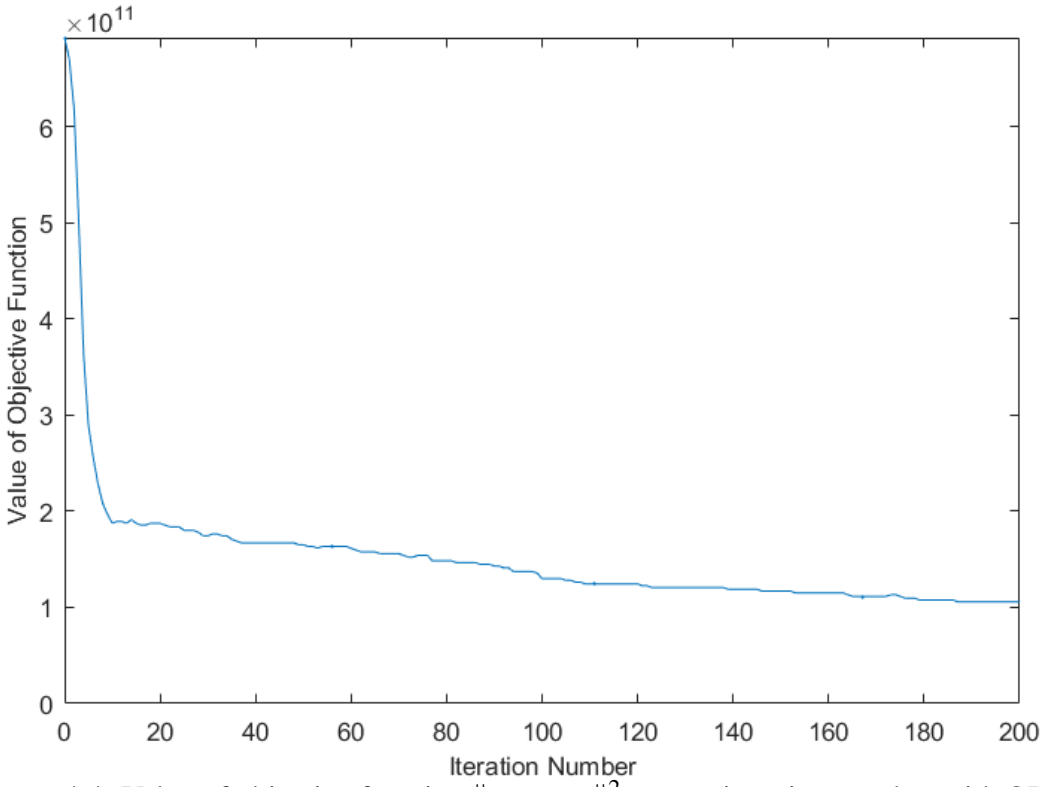


Figure 4.4: Value of objective function $\|WY - X\|_F^2$ versus iteration number with ORL database.

4.6.2 ORL Face Database

ORL face database consists of four hundreds facial-images from 40 males and females. 10 sample images are taken for each person with variances in illumination, expressions and details. Each image is cropped to 92×112 pixels. Images are frontal view of the faces with rotation and tilting tolerance of 20°. The proposed method is tested with feature dimensions of 42, 72, 100 and 168.

Table 4.1: Recognition rate (%) for NN, SVM, SRC and proposed method (STLC) on ORL database for several dimensions (D).

$Dimension(D)$	42	72	100	168
NN	90.6	91.1	90.8	91.00
SVM	89.0	90.7	90.9	92.9
SRC	91.6	93.00	90.8	78.50
$STLC$	**87.2**	**89.8**	**92.0**	**92.5**

As shown in table 4.1, the proposed method achieves a recognition rate of 92% and

Table 4.2: Recognition rate (%) for NN, SVM, SRC and proposed method (M) on extended Yale B database for several dimensions (D).

M \ D	56 8×7	120 12×10	224 16×14	504 24×21	1116 36×31
NN	50.44	62.54	68.86	75.44	78.60
SVM	90.79	92.72	93.86	94.65	94.47
SRC	94.82	95.53	96.93	97.28	95.39
Proposed method	**79.12**	**92.81**	**95.88**	**97.11**	**98.16**

92.5% at feature dimensions of $100D$ and $168D$, respectively. STLC outperforms the other methods at $100D$. At lower dimensions the other methods achieves a slightly higher performance than STLC. In general, with ORL database, the recognition rates for all of these methods are close.

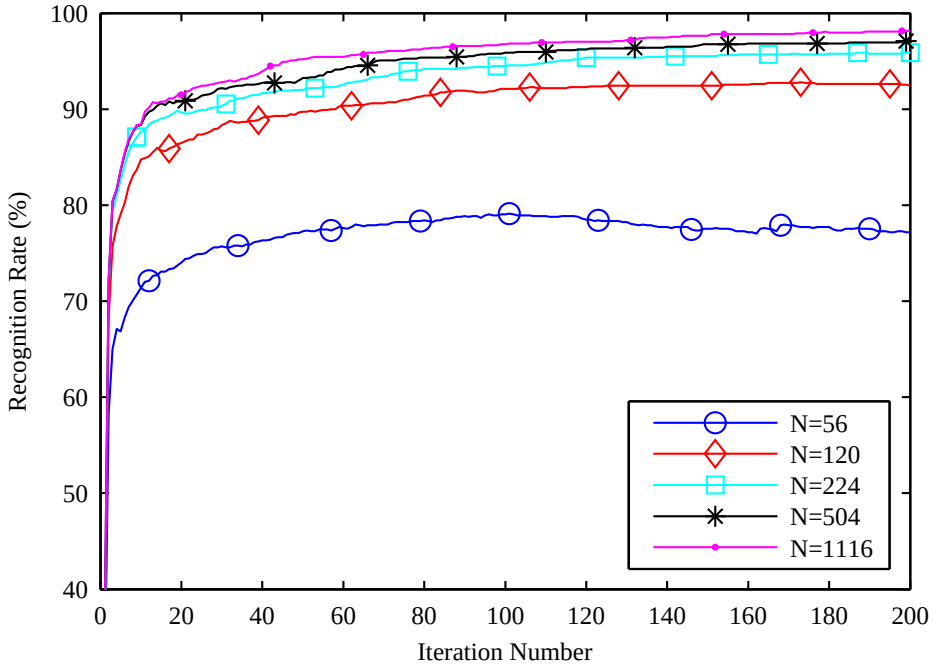


Figure 4.5: Recognition rate vs iteration number with changing dimensions on Extended Yale B database.

4.6.3 Extended Yale B Face Database

This database consist of 38 persons with total of 2414 face images taken from different angles under various lightning conditions [97]. They were cropped to size 192×168 pixels. Each person has about 60 images. 30 images used for training and

30 images for testing. The recognition rates and the computational cost are computed for images down sampled by ratios $1/24$, $1/16$, $1/12$, $1/8$ and $1/5$. The corresponding dimensions are 56, 120, 224, 504 and 1116.

In terms of recognition rate, the algorithm for all dimensions converges after few number of iterations as shown in figure 4.5. It is clear that the recognition rates increase with larger dimensional feature spaces. The difference decreases with large dimensions. Table 4.2 shows the recognition accuracy for the proposed algorithm and other known methods. The proposed method outperforms the NN in all dimensions. SRC and SVM have higher accuracy at low dimensional features. As the dimension increases the proposed method becomes comparable to SRC and SVM until it outperforms them. We note that the recognition accuracy of SRC drops at $1116D$, this is due to the fact that SRC works well for under-determined system of linear equations [58], but with this dimension, the size of the system matrix A in equation 2.8 is 1116×1140.

Figure 4.6 shows the computational time for each dictionary update step because. It has the dominant computational cost. the computational time increases with the dimension of feature space, but still moderate. Once the dictionary constructed, The classification process is very fast and needs very computational cost. The higher dimensionality does not severely affect the computation time of the classification process in the proposed method as it does with SRC and SVM.

Table 4.3 shows the average time elapsed in seconds in the classification stage for each test image in SVM, SRC and STLC. SRC implementation uses Matlab code $\ell_1$eq_pd from $\ell_1$magic [105]. It is obvious from the table that the time required for the

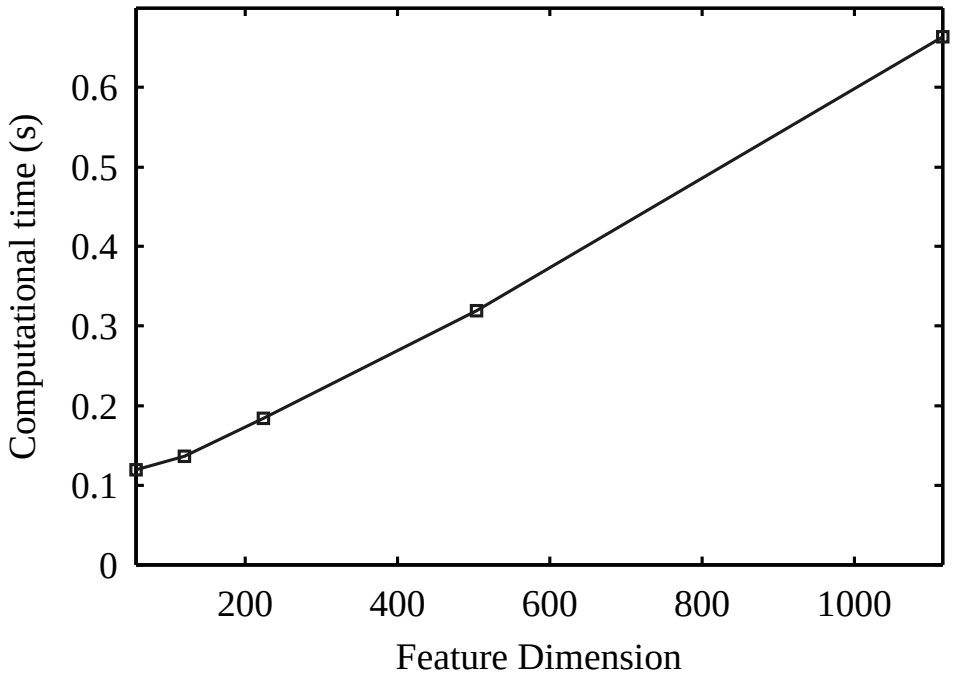


Figure 4.6: Computational time in seconds for every outer iteration vs. dimension

Table 4.3: Average computational time for each test image in seconds for SVM, SRC and proposed method (M) on extended Yale B database for several dimensions.

D / M	56 8×7	120 12×10	224 16×14	504 24×21
SVM	0.8056	0.8346	0.9779	1.1778
SRC	0.8021	1.2575	1.6550	1.5794
Proposed method	**0.0005**	**0.0007**	**0.0010**	**0.0014**

proposed method to identify a new image is about 1000 times less than SRC and SVM at 504D. The difference of the computational times between the proposed method and other methods is due to its simplicity of the classification process which consists of one matrix-vector multiplication and norm comparisons. For SRC method, two observations can be made. The first observation is that SRC has the highest computational time since it needs to perform $\ell_1 - minimization$ for each test image. This comparison shows the advantage of the proposed method, It achieves both goals; the high recognition rate and the fast recognition process.

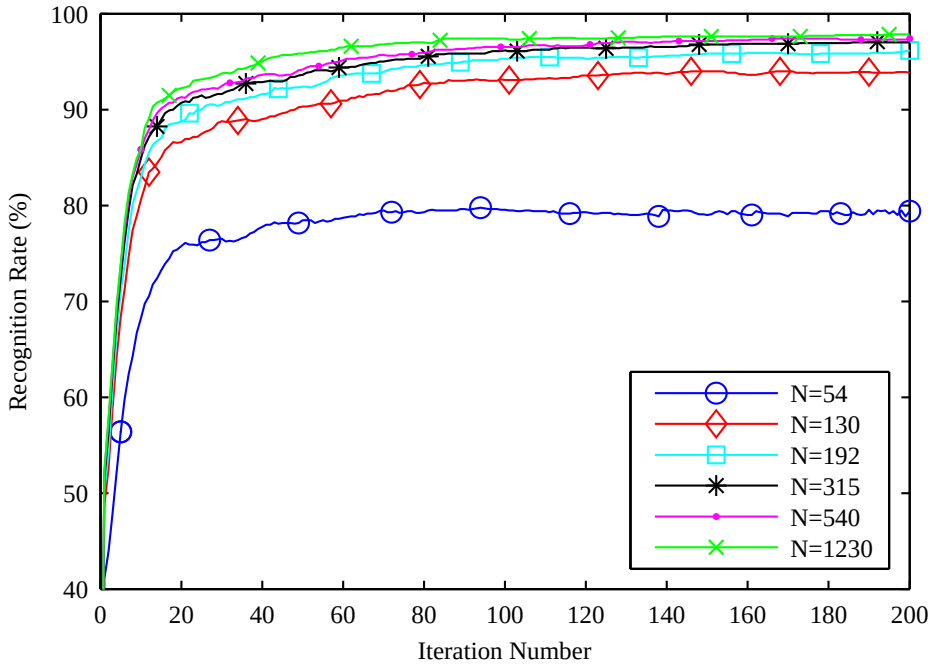


Figure 4.7: Recognition rate vs iteration number with changing dimensions on AR database.

4.6.4 AR Database

AR has more than 4000 images for 126 males and females. For each person, 26 face images taken under various conditions. each image is cropped to dimension 165×120. In our test, We randomly select 100 persons for 50 men and 50 women. For each individual, 26 images, half of them are selected randomly for training and the others for testing. After converting the images to gray scale, the recognition rates are computed for feature dimensions of 54, 130, 192, 315, 540 and 1230, corresponding to down-sampling by ratios $1/18$, $1/12$, $1/10$, $1/8$, $1/6$ and $1/4$.

As shown in figure 4.7, the convergence speed with AR database is almost the same as the convergence speed with Yale database. The recognition rate increases rapidly at the first few iterations.

Figure 4.8 shows the recognition rates for the proposed method, NN, SVM and SRC. STLC achieves recognition accuracy between 94.0% for $130D$ feature space and

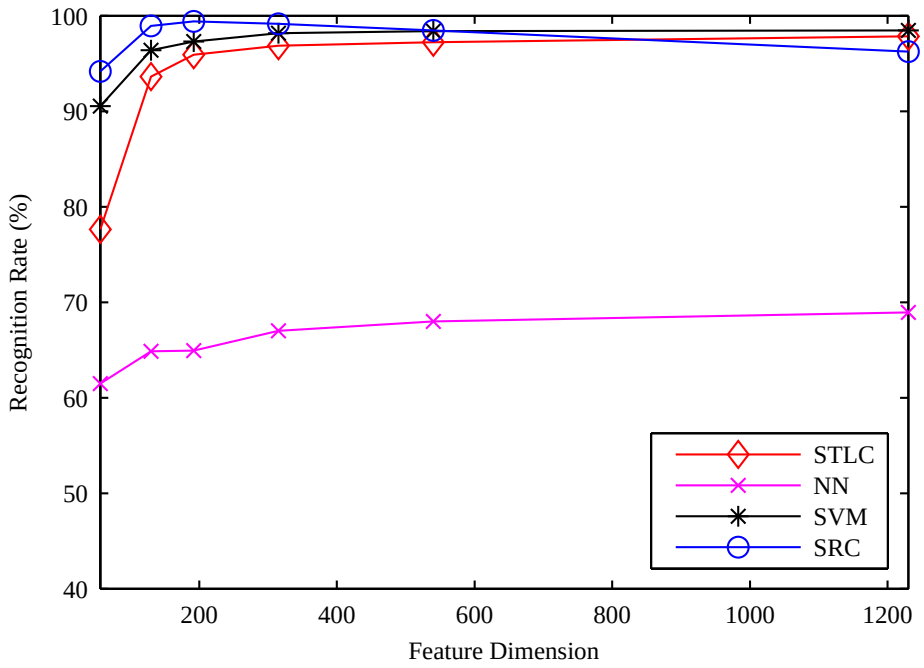


Figure 4.8: Recognition rate for NN, SVM, SRC and STLC methods on AR database for several dimensions.

97.4% for 1230D feature space. The comparison among these methods is the same as in extended yale B database.

4.7 Conclusion

We proposed a new face recognition algorithm, namely STLC. The proposed method learns a transformation dictionary to sparsify the image in a discriminative pattern such that the nonzero sparse coefficients are in places related to their class. Unlike other sparsity based methods, the proposed method has a very fast classification process that only compute and compare the norm of the coefficients belonging to each class. Simulation results on two different databases have shown the advantage and the high accuracy of this approach. Although other techniques have shown to be better in terms of accuracy at low dimensions, STLC still distinguished with its low computational cost and can be used as a stand-alone classifier.

Chapter 5

SPARSE ℓ_2-NORM REGULARIZED REGRESSION FOR FACE CLASSIFICATION

5.1 Introduction

In this chapter a second new idea for face recognition is proposed. This method is a regression based method. It transforms the images into sparse vector using regularization terms.

The contributions of this chapter can be stated as follows:

- A regression based image classification algorithm is proposed. The method uses $\ell_2 -$ norm regularized objective function to prevent overlapping of nonzero coefficients that belong to different classes.
- The proposed method transform images to sparse vectors using $\ell_0 -$ norm constraint.
- Simulation results verify that our method is competitive and superior to the alternative projection based methods.

The remaining sections of this chapter are arranged as following. Section 5.2 discusses the details of the method. Section 5.3 presents the solution procedure for the method. Classification procedure is explained in section 5.4. Simulations and experiments are discussed in section 5.5. Finally, section 5.6 summarizes and concludes the proposed

method.

5.2 Problem Formulation and Objective Function

Given n training face images for K persons. In matrix form, the n_k images from k^{th} class can be represented as:

$$Y_k = [y_{k1}, y_{k2}, \cdots, y_{kn_k}] \in \mathbb{R}^{N \times n_k}, \;\; k = 1, 2, ..., K \tag{5.1}$$

And the matrix representation of all subjects is represented as:

$$Y = [Y_1, Y_2, \cdots, Y_K] \in \mathbb{R}^{N \times n} \tag{5.2}$$

To train the dictionary W, we have formulated the following problem

$$\begin{aligned} \text{(P1)} \min_{W,X} \quad & \|WY - X\|_F^2 - \lambda \log \Big[\sum_{\substack{i=1 \\ i \neq j}}^{K} \sum_{j=1}^{K} \|W_i Y_i - W_j Y_j\|_F^2 \Big] \\ \text{s.t.} \quad & \|X_i\|_0 \leq s_i \; \forall i \end{aligned} \tag{5.3}$$

where W_i is a sub-matrix of W such that $W = [W_1^T, W_2^T, \cdots, W_K^T]^T$. For simplicity, we assume here that each submatrix W_i is a representation of consecutive number of rows in W. But in the real implementation of the algorithm, W_i represent distributed rows in W where we keep the largest s coefficients in each column as explained in the next section.

The rational of this objective function is to increase the inner product between the transformations of the images that belong to the same class and to decrease the inner product between the transformations of the images that belong to different class. This claim becomes obvious when we expand second term in the objective function as

$$\|W_i Y_i - W_j Y_j\|_F^2 = \|W_i Y_i\|_F^2 + \|W_j Y_j\|_F^2 - 2(W_i Y_i)^T (W_i Y_i) \tag{5.4}$$

In other words, ideally, W will transform images to sparse vectors where the positions of nonzero coefficients for image transformation will not overlap with image transformation of other classes.

5.3 Solution Procedure

The solution of the nonconvex optimization problem (P1) can be achieved iteratively by alternating between two steps; sparse coding and dictionary update as shown in algorithm 1 .

Algorithm 1 solution Procedure

INPUT: Training images: $Y_k \in \mathbb{R}^{N \times n_k},\ \ k = 1, 2, ..., K$

1: Initialize zero matrix $Index$.
2: Initialize zero matrix X.
3: **for** $i = 1$ to $Number\ of\ iterations$ **do**
4: $X = WY$-Transformation of class k training images
5: $X = f(x)$- where f is a function that keeps the largest s coefficients in each column of X and zeroing all others.
6: **for** $i = k$ to K **do**
7: $Index[:, k]$= vector of indices of the largest s coefficients in for each class after excluding indices for previous classes $Index[:, j], j < k$
8: Calculate W_k according to equation 5.16

5.3.1 Sparse Coding

Solve (P1) by updating the vector X while keeping the dictionary W fixed. The given problem turns to the following equation

$$\min_X \|WY - X\|_F^2,\ s.t.\ \|X_i\|_0 \leq s\ \forall i \tag{5.5}$$

The solution of the above equation for X can be found by zeroing all except the largest s coefficients in each column of matrix WY. The indices of the remaining nonzero coefficients are kept in a database to be used in image classification process.

5.3.2 Dictionary Update

Solve (P1) by updating W while keeping X fixed. In this step (P1) is switched to following minimization problem

$$\min_{W} \quad \|WY - X\|_F^2 - \lambda \log \Big[\sum_{\substack{i=1 \\ i \neq j}}^{K} \sum_{j=1}^{K} \|W_i Y_i - W_j Y_j\|_F^2 \Big] \tag{5.6}$$

Optimization problem (5.6) has a convex differentiable objective function in W. Minimizing this objective function can be achieved by computing its derivative with respect to W and then solve for W that makes the gradient zero. The gradient of the first part is

$$\nabla_W \|WY - X\|_F^2 = 2WYY^T - 2XY^T \tag{5.7}$$

To find the derivative of the second part of the objective function, consider

$$g(W) = \sum_{\substack{i=1 \\ i \neq j}}^{K} \sum_{j=1}^{K} \|W_i Y_i - W_j Y_j\|_F^2 \tag{5.8}$$

and

$$f(W) = \log \big[g(W) \big] \tag{5.9}$$

Now, the derivative of $f(W)$ can be computed using the partial derivative $\frac{\partial g}{\partial W_k}$ as follows

$$\nabla_W f = \frac{1}{g(W)} \Big[\frac{\partial g}{\partial W_1}, \frac{\partial g}{\partial W_2}, \ldots, \frac{\partial g}{\partial W_K} \Big]^T \tag{5.10}$$

To compute $\frac{\partial g}{\partial W_k}$, We first expand (5.8) to find the terms that includes W_k

$$g = \sum_{\substack{j=1 \\ i \neq j}}^{K} \|W_k Y_k - W_j Y_j\|_F^2 + \sum_{\substack{i=1 \\ i \neq j}}^{K} \|W_i Y_i - W_k Y_k\|_F^2 \tag{5.11}$$

$$+ \sum_{\substack{i=1 \\ i \neq k \\ i \neq j}}^{K} \sum_{\substack{j=1 \\ j \neq k}}^{K} \|W_i Y_i - W_j Y_j\|_F^2 \tag{5.12}$$

$$= 2\sum_{\substack{j=1 \\ j\neq k}}^{K} \|W_k Y_k - W_j Y_j\|_F^2 + \sum_{\substack{i=1 \\ i\neq k \\ i\neq j}}^{K} \sum_{\substack{j=1 \\ j\neq k}}^{K} \|W_i Y_i - W_j Y_j\|_F^2 \tag{5.13}$$

The second part of this equation does not depend on W_k, so its derivative is zero. Hence

$$\begin{aligned} \frac{\partial g}{\partial w_k} &= 4\sum_{\substack{j=1 \\ j\neq k}}^{K} (W_k Y_k - W_j Y_j) Y_K^T \\ &= 4\big[(K-1)W_k Y_k - \sum_{\substack{j=1 \\ j\neq k}}^{K} W_j Y_j\big] Y_K^T \end{aligned} \tag{5.14}$$

To find the minimum value of the objective function we solve the following equation for every W_k

$$W_k\big[2YY^T - 4\lambda(K-1)Y_k Y_K^T\big] - 2XY^T + 4\lambda\Big(\sum_{\substack{j=1 \\ j\neq k}}^{K} W_j Y_j\Big) Y_K^T = 0 \tag{5.15}$$

Finally, we obtain the following closed form solution for each W_k

$$W_k = \Big[2YY^T - 4\lambda(K-1)Y_k Y_K^T\Big]^{-1}\Big[2XY^T - 4\lambda\Big(\sum_{\substack{j=1 \\ j\neq k}}^{K} W_j Y_j\Big) Y_K^T\Big] \tag{5.16}$$

5.4 Classification

Given a new face image y_{new} to be classified to one of the classes, a new test image will be transformed using the dictionary W as

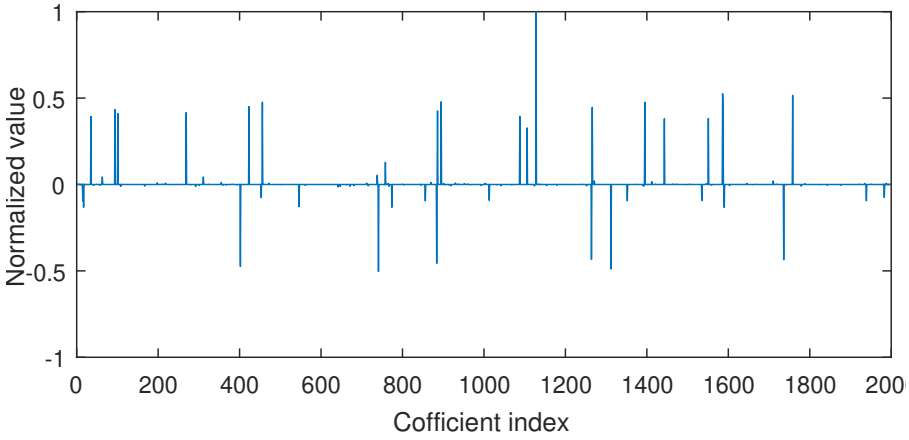
$$x = W y_{new} \tag{5.17}$$

Consider a function $\delta_k(x) : \mathbb{R}^L \to \mathbb{R}^s$ for each class k, which only selects the coefficients in x that correspond to class k. Then based on the maximum ℓ_2-norm of the selected coefficients, the test image y will be classified to its class as

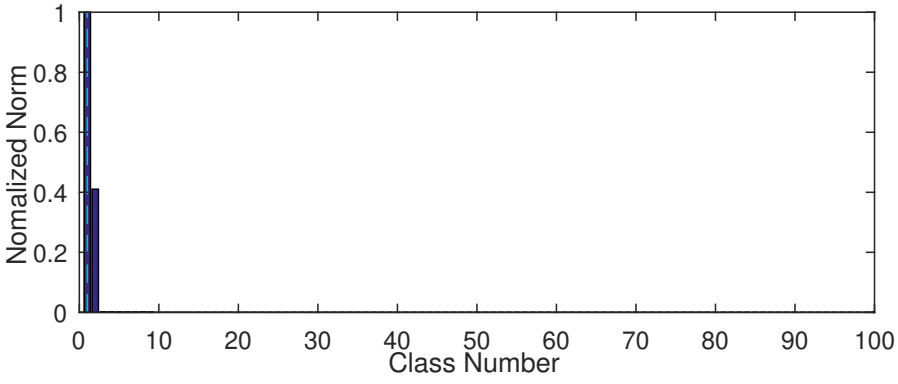
$$class(y_{new}) = \underset{k}{\operatorname{argmax}} \|\delta_k(x)\|_2^2 \tag{5.18}$$

Figure 5.1 shows an example of image transformation. In this example, the test image in figure 5.1(a) is selected from the first class of AR database. The normalized sparse vector x that is obtained is shown in figure 5.1(b). Figure 5.1(c) shows the normalized norms of coefficients of vector x for each class, it is obvious that the first class has the maximum norm.

(a)



(b)



(c)

Figure 5.1: Example of an image transformation. (a) Test image. (b) The sparse coefficient vector x. (c) Norm of coefficients for each subject.

5.5 Experimental Validation

We test our algorithm with the standard face databases ORL [93], AR [95], the extended-YaleB face database [94] and LFW databases [96]. The performance of the proposed algorithm was compared with other projection based methods in literature, i.e., PCA, LFDA [6], LPP [7], SPP [106], CRP, and SVM [62]. The classifier that is used with these algorithms is k-nearest neighbor.

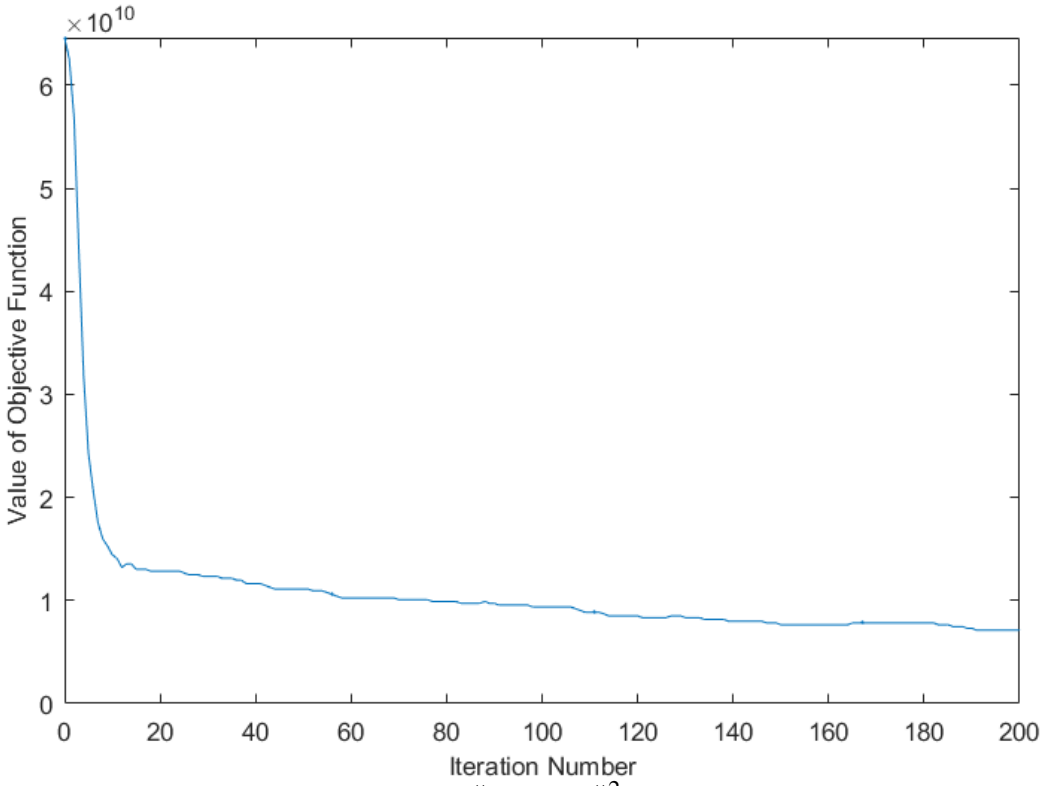


Figure 5.2: Value of objective function $\|WY - X\|_F^2$ versus iteration number with ORL database.

In the experiments on LFW database, 10-fold validation is used. On other databases, l face images for each subject are randomly selected to construct the transformation dictionary and the remaining are used for testing. Number of training images l={2, 6} for ORL face database, l={4, 8, 16} for AR face database and l={8, 16} for extended YaleB face database.

The regularization variable λ is set to 1×10^{-9} . Number of the iterations is 50. The ratio of nonzero coefficients is selected to be 10% of the total coefficients . The simulations are carried out with Intel i77500U CPU at 2.7GHz and 2.9 GHz and 12GB memory.

5.5.1 Stability and Convergence of The Proposed Method

The stability and convergence of the proposed method is tested with ORL Database. As shown in figure 5.2, the value of the objective is monotonically decreasing and reach the stability after few iterations.

5.5.2 ORL Database

ORL [93] contains 400 facial-images captured for 40 persons, 10 sample images for each one. They were taken with variances in illumination, facial expressions and facial details. Each image was cropped to size 92×112 pixels. Simulation results listed in table 5.1 shows that our method achieves the highest recognition rate.

Table 5.1: Recognition rates (%) of state-of-the-art and the proposed methods on ORL face database

Training samples	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
l=2	74.6	92.6	88.5	90.1	89.8	92.4	93.1
l=6	82.4	97.8	95.7	97.4	96.6	97.8	98.3

5.5.3 AR Database

A subset of AR face database has been randomly selected. The selected subset consists of 100 persons for 50 women and 50 men. For each one, 26 gray scale face images of dimension 165×120 pixels are used for training and testing the proposed algorithm. Sample face images are shown in figure 5.3.

Simulation results are shown in table 5.2. We can conclude that the proposed method achieves the best rates among all methods. The recognition rates for the proposed

method and CRP are very close when the number of training images is 16.

Figure 5.3: Sample images of AR database.

Table 5.2: Recognition rates (%) of state-of-the-art and the proposed methods on AR face database.

Training samples	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
l=4	55.6	69.7	67.1	66.2	74.3	70.4	75.1
l=8	61.7	78.6	69.4	79.6	79.5	76.2	81.4
l=16	80.7	94.7	94.4	97.2	98.0	94.9	98.1

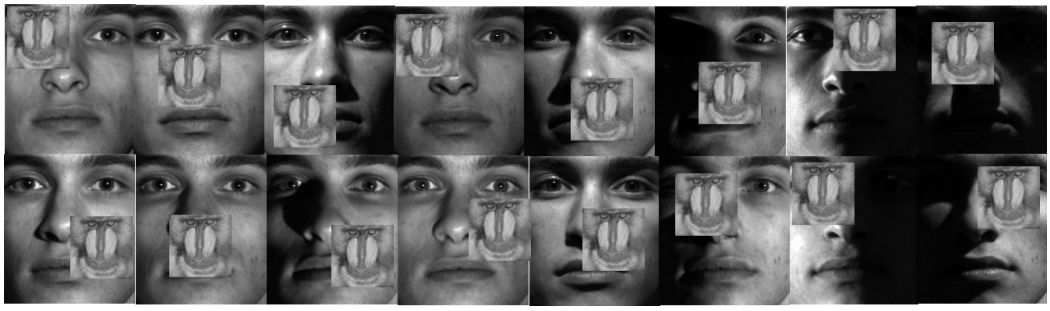
5.5.4 Extended Yale B Database

Extended Yale B [94] comprises 38 persons with 2414 frontal person's face images. They were captured from different angles under different lightning conditions. The images were cropped to the size 192×168 pixels. Each person has about 60 samples. 30 image samples are used for training and 30 image samples for testing. In these experiments, PCA was used to reduce the feature dimensions to 120. Figure 5.4(a) shows subset of face images of one person. Table 5.3 shows the simulation results for different number of randomly selected training images.

The proposed method is also tested under occlusion and corruption as shown in figure 5.4(b) and figure 5.4(c), respectively. Table 5.4 shows the recognition rates under 20% occlusion, where baboon image is added to the original images, as shown in figure 5.4(b). Results of experiments under 20% corruption are shown in table 5.5.

All simulation results in tables 5.3, 5.4 and 5.5 show the superiority of the proposed

(a)



(b)

(c)

Figure 5.4: Sample images of Extended Yale B database of (a) Samples for one person. (b) Samples with 20% occlusion . (c) Samples with 20% corruption.

Table 5.3: Recognition rates (%) of state-of-the-art and the proposed methods on YaleB database

Training samples	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
l=8	63.6	78.2	70.3	81.4	80.6	79.1	82.2
l=16	72.3	95.4	95.3	97.2	96.3	95.6	97.3

method over other methods. The proposed method achieves 97.3% recognition rate at $l = 16$. Even though the performance degrade to 90.7% in case of occluded and to 91.6% in case of corrupted images, The proposed method still have the highest recognition rate compared to other methods under the same conditions.

Table 5.4: Recognition rates (%) of state-of-the-art and the proposed methods on Extended YaleB face database with 20% block occlusion

Training samples	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
l=8	52.4	60.7	69.4	69.8	71.6	61.3	71.1
l=16	60.2	85.9	86.6	82.3	90.4	87.6	90.7

Table 5.5: Recognition rates (%) of state-of-the-art and the proposed methods on Extended YaleB face database with 20% corruption

Training samples	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
l=8	53.6	73.8	71.8	72.8	72.3	73.6	73.1
l=16	64.3	84.1	83.9	82.3	91.1	86.4	91.6

5.5.5 Experiments on LFW Database

For this database, 100 subjects of LFWa database [107] were used. For each person, 6 images are selected. The images are cropped to eliminate the background, and resized to 64×64. Figures 5.5(a, b, c) show samples of LFW, occluded LFW, and corrupted LFW database, respectively .

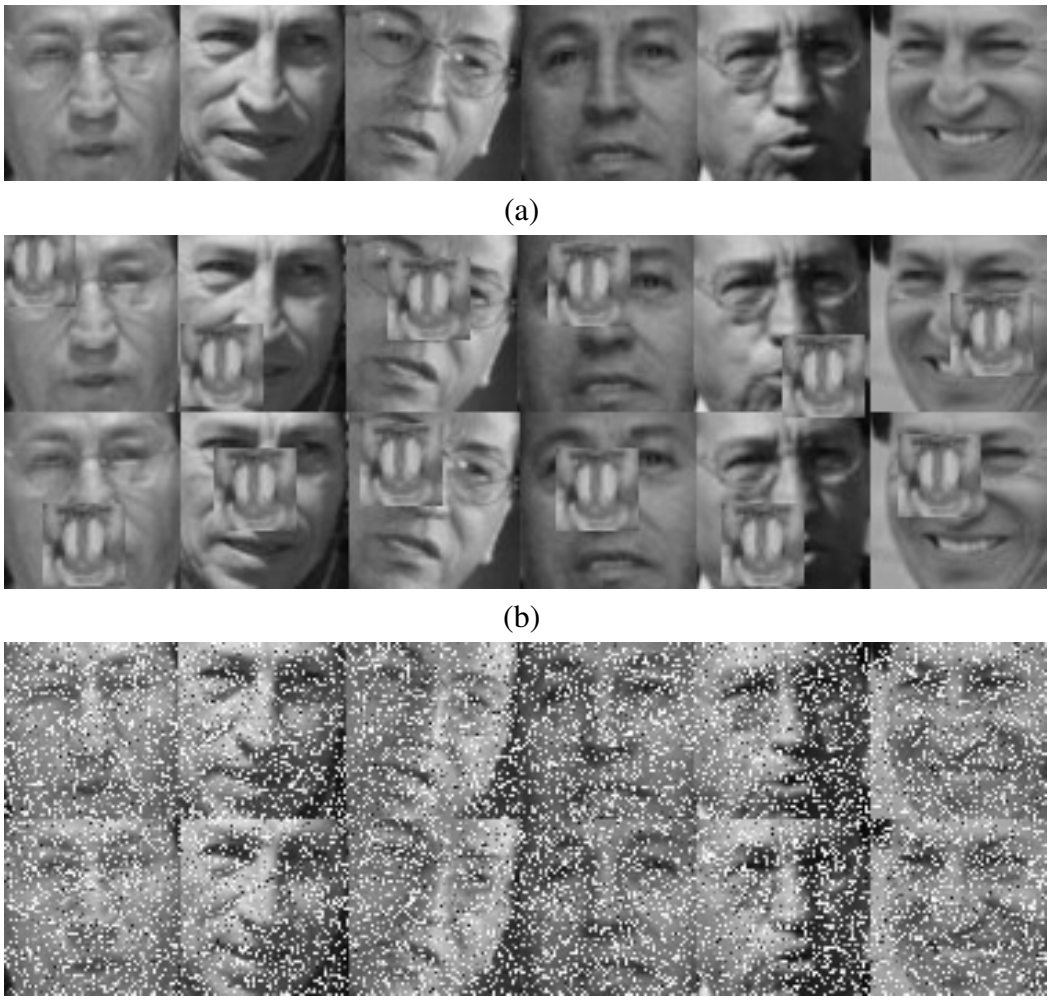
Table 5.6: Recognition rates (%) of state-of-the-art and the proposed methods on LFW database

Database	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
LFW	66.9	94.1	92.5	93.8	95.1	94.2	95.3
Occluded LFW	60.9	86.8	81.3	86.1	83.2	86.5	89.1
Corrupted LFW	61.5	88.7	85.8	87.4	84.1	89.1	89.4

As in experiments with the previous databases, our method gets the highest recognition rates among all methods with LFW database as shown in table 5.6.

5.6 Conclusion

A new regression based face recognition algorithm is proposed. The method uses ℓ_0-norm to transform the image into a sparse vector. It uses ℓ_2-norm regularization to prevent the overlapping of nonzero coefficients that belongs to different subjects. The proposed method is tested with different face databases. It is compared with other well-



(a) (b) (c)

Figure 5.5: Sample images of LFW database of (a) Samples for one person. (b) Samples with 20% occlusion . (c) Samples with 20% corruption.

known methods. Results show the superiority of accuracy of our method. They also show the robustness of the method under occlusion and corruption. Another advantage of the proposed method is the low computational cost, since it only contains matrix vector multiplication and norm computation to achieve the classification task.

Chapter 6

SPARSE REGULARIZED REGRESSION BASED METHOD FOR FACE RECOGNITION

6.1 Proposed Method

In this chapter, a novel $\ell_2 -$ norm based regression method for feature extraction and face images classification is proposed. The proposed algorithm consists of two stages. The first stage is the projection matrix learning stage, where the training images are used to construct a projection matrix such that it can transform the images into sparse vectors. The sparsity patterns of the vectors are different for images of different classes. The second stage is classification stage in which the algorithm uses the matrix to transform the new images into sparse vectors, and then classify them based on the obtained sparsity pattern.

The remaining parts of this chapter are organized in the following way: Section 6.1 provides detailed explanations of the proposed method. Section 6.2 presents the solution procedure for the method. Classification process is explained in section 6.3. In section 6.4, simulation results are presented and discussed. Finally, summary and conclusions are presented in section 6.5.

6.1.1 Problem Formulation

Given n total number of training face images for K different persons, assume that the k^{th} person has n_k different images, we represent the face images of k^{th} person in a

matrix form as

$$Y_k = [y_{k1}, y_{k2}, \cdots, y_{kn_k}] \in \mathbb{R}^{N \times n_k}, \;\; k = 1, 2, ..., K \tag{6.1}$$

where N represent the dimension of each image. The whole training images for all persons also are arranged in matrix form as

$$Y = [Y_1, Y_2, \cdots, Y_K] \in \mathbb{R}^{N \times n} \tag{6.2}$$

6.1.2 Projection Matrix Learning

We need to train a transformation dictionary W that transforms an image to a sparse vector. The sparsity patterns of the vectors have to be different for images that belong to different classes in order to be used in image classification process. The following optimization problem has been formulated to train the matrix W

$$\begin{aligned} (\text{P1}) \underset{W,X}{\text{argmin}} \quad & \|WY - X\|_F^2 + \lambda_1 \sum_{i=1}^{K} \|WY_i - W\overline{Y_i}\mathbf{1}_{\mathbf{n_i}}^{\mathbf{T}}\|_F^2 \\ & - \lambda_2 \log \Big[\sum_{\substack{i=1 \\ i \neq j}}^{K} \sum_{j=1}^{K} \|WY_i - WY_j\] \|_F^2 \\ \text{s.t.} \quad & \|\overline{X_i}\|_0 \leq s_i \; \forall i \end{aligned} \tag{6.3}$$

where λ_1 and λ_2 are the regularization parameters. $\mathbf{1}_{\mathbf{n_i}}^{\mathbf{T}} \in \mathbb{R}^{1 \times n_i}$ is row vector of ones used to replicate the mean vector. Columns of X represent the sparse transformation of training images. $\overline{Y_i}$ is the mean of class i images.

$$\overline{Y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \tag{6.4}$$

The goal of the proposed optimization problem is to train a transformation matrix that

can transform the images to sparse codes to be used for classification. The image transformation x has the following properties

- It is a sparse vector with sparsity level s
- Nonzero coefficients are at the same locations for images of the same class and do not overlap with nonzero coefficients with transformations of images that belong to different classes.
- The ℓ_2-norm of the nonzero coefficients should be maximized.
- The similarity between image transformations of different classes should be minimized.

These properties are guaranteed by the optimization problem as explained next. Determining the positions of nonzero coefficients will be explained in the solution procedure of the optimization problem.

6.1.3 Rationale of The Objective Function

The term $\|WY - X\|_F^2$ represent the error between the ideal sparse transformation X and the actual image transformation WY. This term enforces the image transformation to be sparse and close to the code X, where X is the sparse code that is obtained by preserving the s largest coefficients in each column of matrix WY.

The second term of the proposed objective function, $\sum_{i=1}^{K} \|WY_i - W\overline{Y_i}\mathbf{1}_{\mathbf{n_i}}^{\mathbf{T}}\|_F^2$, enforces the transformations of images of the same class to be close to a specific central point which is the transformation of the mean vector $\overline{Y_i}$. This will increase the similarity between image transformations of the same class as shown in figure 6.1.

The term $-\log\left[\sum_{i=1}^{K}\sum_{\substack{j=1 \\ i\neq j}}^{K} \|WY_i - WY_j\]\|_F^2\right]$ maximizes the ℓ_2-norm of image

transformation vectors while reducing the similarity between image transformations of different classes. This explanation becomes clear when we expand the function inside the log function as

$$\|WY_i - WY_j\|_F^2 = \|WY_i\|_F^2 + \|WY_j\|_F^2 - 2(WY_i)^T(WY_j) \tag{6.5}$$

Ideally WY_i and WY_j will be orthogonal to each other.

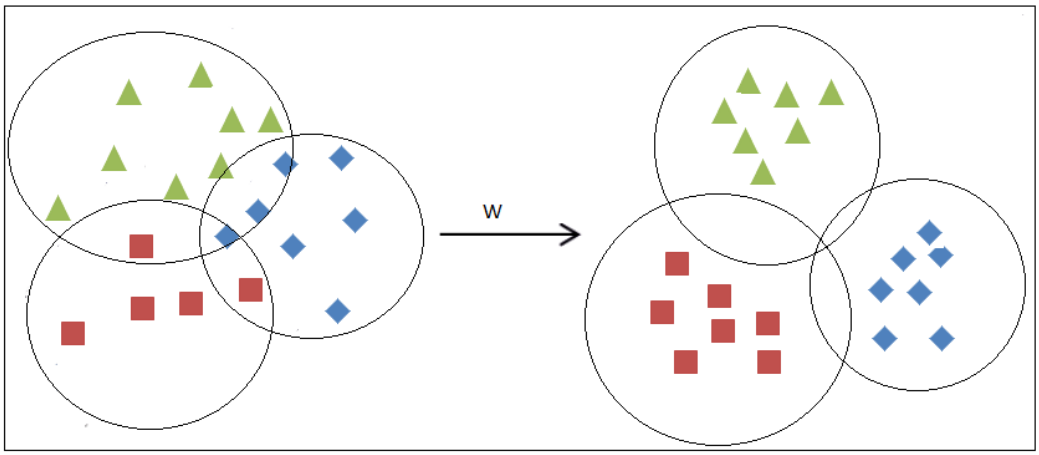


Figure 6.1: Effect of transformation on distances between data points of different classes

6.2 Solution Procedure

The proposed optimization problem (P1) can be solved for transformation matrix W and the sparse code$X$ by iteratively updating X and W.

6.2.1 Update X

Keep the transformation matrix W fixed, update the sparse code matrix X. In this step the proposed problem becomes

$$\underset{X}{\operatorname{argmin}} \|WY - X\|_F^2, \; s.t. \; \|X_i\|_0 \leq s \; \forall i \tag{6.6}$$

As shown in algorithm 2, this optimization problem can be solved for X by finding the

Algorithm 2 Calculate X

INPUT: Dictionary: $W \in \mathbb{R}^{L\times N}$. Training images: $Y_k \in \mathbb{R}^{N\times n_k}$, $k = 1,2,...,K$

1: Initialize zero matrix $Index$.
2: Initialize zero matrix X.
3: **for** $k = 1$ to K **do**
4: $T_k = WY_k$-Transformation of class k training images
5: $m_k = \frac{1}{n_k}\sum_{j=1}^{n_k} T_{kj}$-mean vector of class k transformation
6: $Index[:,k]$= vector of indices of the largest s coefficients in m_k after excluding indices for previous classes $Index[:,j], j < k$
7: **for** $i = 1$ to n_k **do**
8: $X_k[Index[:,k],i] = T_k[Index[:,k],i], i = 1,...,n_k$, keep the largest s coefficients
9: $X = [X_1,X_2,...,X_K]$

positions of the s largest coefficients in the average vectors for images transformation of each class k.

$$m_k = \frac{1}{n_k}\sum_{j=1}^{n_k} WY_{kj}, \quad K = 1,...,K \tag{6.7}$$

Then zeroing all coefficients in X, except those in the positions of the largest s coefficients in the mean vector of the class m_k. Additional condition should be satisfied, which is the nonzero coefficient positions for each class are mutually exclusive sets.

6.2.2 Update W

Keep the sparse code matrix X fixed. Update the transformation dictionary W. The optimization problem becomes

$$\begin{aligned}\underset{W}{\operatorname{argmin}} \quad & \|WY - X\|_F^2 + \sum_{i=1}^{K} \|W\left(Y_i - \overline{Y_i}\right)\|_F^2 \\ & - \lambda \log\Big[\sum_{\substack{i=1\\ i\neq j}}^{K}\sum_{j=1}^{K} \|WY_i - WY_j\Big]\|_F^2\end{aligned} \tag{6.8}$$

Objective function in optimization problem (6.8) is convex differentiable function in W. Gradient descent algorithm [103] can be used to minimize this objective function.

6.3 Classification Procedure

The Algorithm is proposed to recognize and classify a new test face image y_{new} into one of the given subjects. First, The test image is transformed by dictionary W

$$x = Wy_{new} \tag{6.9}$$

Second, find the coefficients $\delta_k(x)$ for each class k. Vector $\delta_k(x)$ is constructed by selecting the coefficients in vector x that correspond to class k. Then, the decision is made by finding the maximum $\ell_2 - norm$ of the class coefficients, the image y_{new} will be classified as

$$class(y_{new}) = \underset{k}{\operatorname{argmax}} \|\delta_k(x)\|_2^2 \tag{6.10}$$

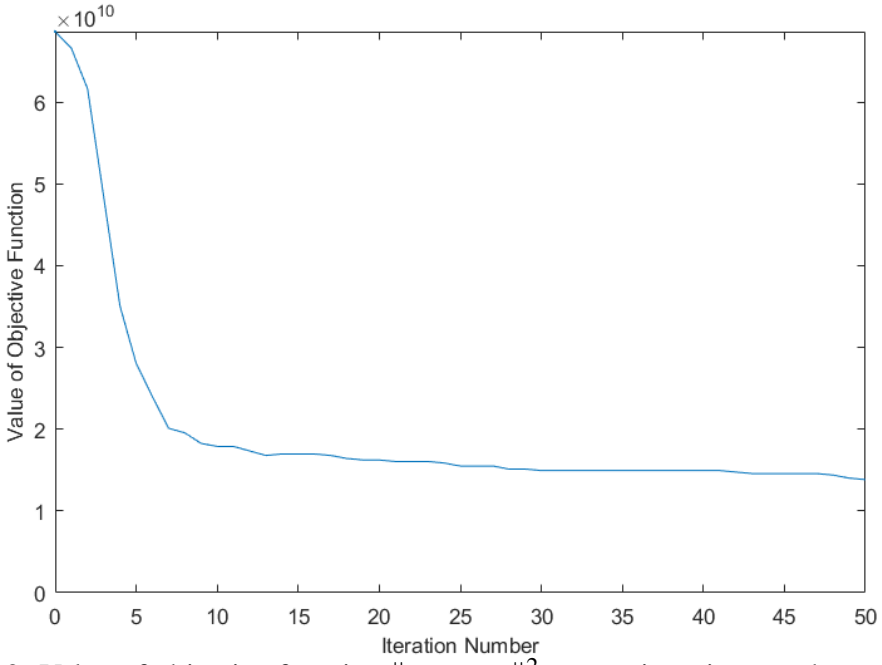


Figure 6.2: Value of objective function $\|WY - X\|_F^2$ versus iteration number with ORL database.

6.4 Experimental Validation

Simulations are executed with different benchmark face databases; ORL [93], the extended-YaleB face database [94], AR [95] and LFW databases [96].

The simulation results obtained for the proposed algorithm are compared with other feature extraction methods in literature. The comparisons are made with PCA, LFDA [6], LPP [7], SPP [106], CRP [62], and SVM. With these methods, the we use k-Nearest Neighbor method.

Regularization parameter λ_1 and λ_2 are set to be 1×10^{-5} . Number of iterations is set to 50. The projection matrix W is randomly initialized. The number of nonzero coefficients in the sparse code is set to be 10% of the total number of coefficients.

6.4.1 Stability and Convergence of The Proposed Method

The stability and convergence of this algorithm is tested with ORL Database. Figure 6.2 shows the progress of the optimization function with iterations. The figures shows that the value of the objective is monotonically decreasing and reach the stability after few iterations.

6.4.2 ORL Database

A database [93] of 400 human face images for forty persons. for each person, 10 images are captured under various conditions with different facial details, facial expressions and illumination conditions. The simulations are performed with l=$\{2, 4, 5\}$ randomly selected training images. The simulation results are shown in Figure 6.3. It shows that the our method has the best performance for all number of training images.

6.4.3 AR Database

A subset of 100 persons in AR database is randomly selected. The subset consists of 50 men and 50 women with 26 gray scale images of size 165×120 pixels for each one.

The proposed method is tested with diverse number of training images. Simulation resultsare shown in figure 6.4. The recognition rates for the proposed method and CRP

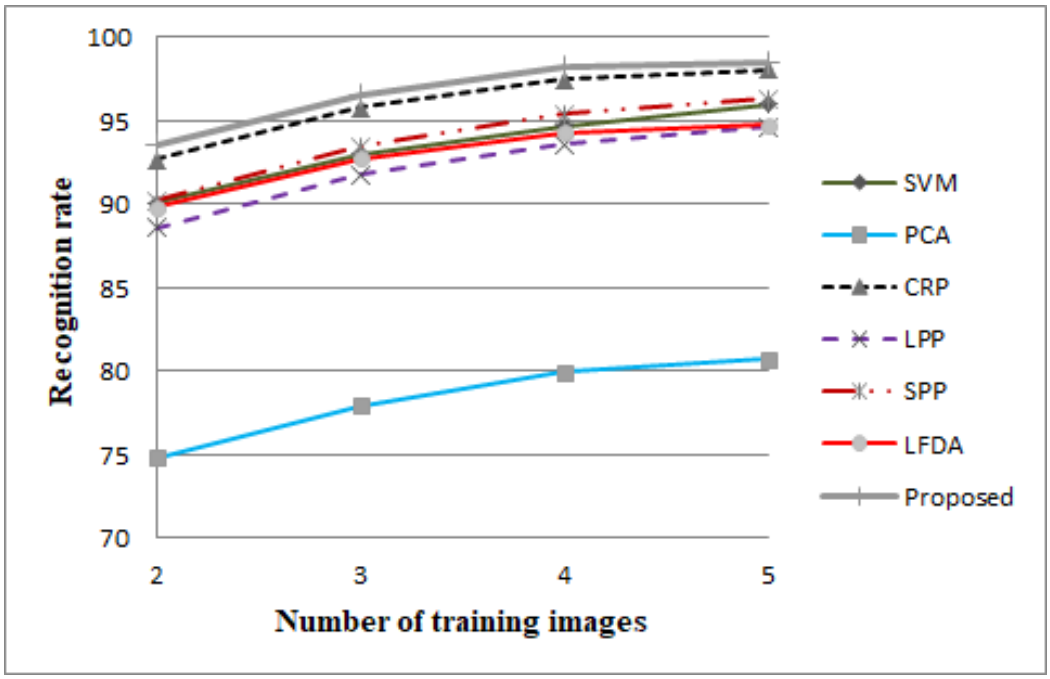


Figure 6.3: Recognition rates (%) of state-of-the-art and the proposed methods on ORL face database versus number of training images

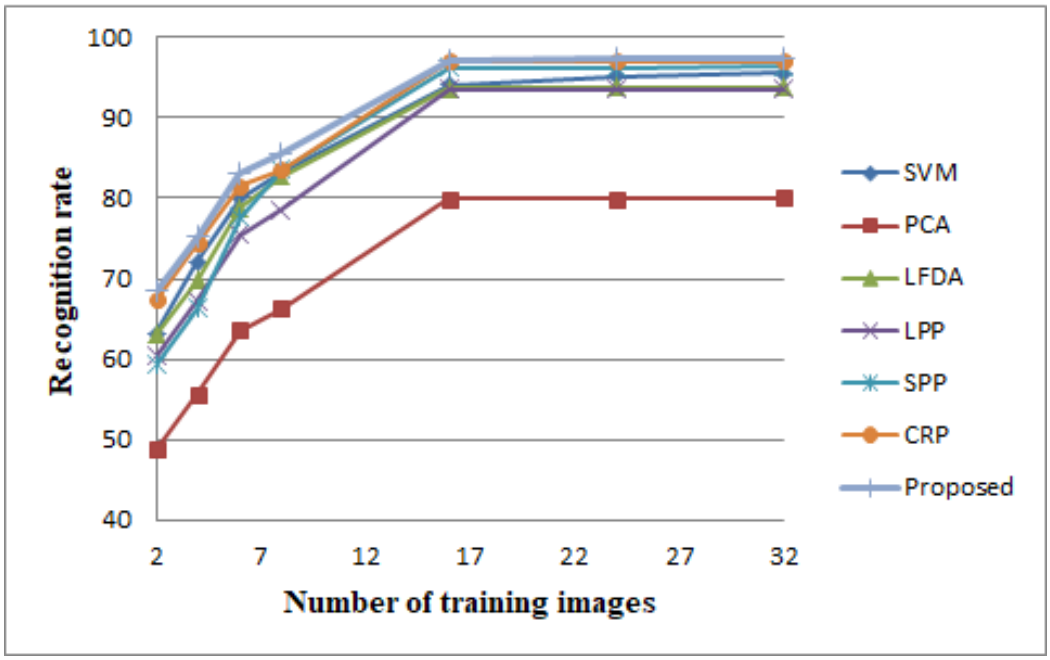


Figure 6.4: Recognition rates (%) of state-of-the-art and the proposed methods on AR face database versus number of training images

are close.

Table 6.1: Recognition rates (%) of state-of-the-art and the proposed methods on Yale B database

Training samples	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
l=2	45.14	59.05	56.41	55.48	60.2	58.3	65.44
l=4	53.94	67.85	65.23	64.26	72.64	68.4	73.99
l=8	63.76	78.34	70.48	81.51	80.72	78.1	82.88
l=16	72.31	95.42	95.41	97.29	96.48	96.3	97.41

Figure 6.5: Samples with corruption in Extended Yale B database

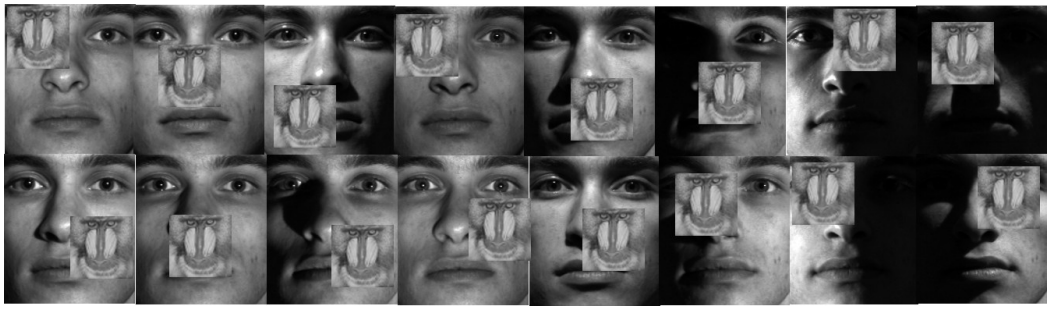


Figure 6.6: Samples with occlusion in Yale database

6.4.4 Extended Yale B Face Database

Extended Yale B [94] is database of 2414 frontal human face images for 38 persons, 60 images for each one. each image dimension is 192×168 pixels. Table 6.1 shows simulations on this database with l=$\{2, 4, 8, 16\}$ randomly selected training images for each person. From table 6.1, we can notice that our method achieved the highest rates.

We also test the method with random block occlusion and corruption. Figures 6.5 and 6.6 show samples images with corruption and occlusion, respectively. In random block occlusion, 20% of the image pixels are covered by baboon image at random locations.

In random corruption case, 20% of the image pixels are corrupted with random noise. Simulation results on Yale database with occlusion and corruption are shown in tables 6.2 and 6.3, respectively. Results clearly reveal the superiority of our method over the others.

Table 6.2: Recognition rates (%) of state-of-the-art and the proposed methods on Extended YaleB face database with 20% block occlusion

Training samples	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
l=2	36.5	50.1	47.6	46.6	54.9	52.1	56.8
l=4	47.1	61.1	58.2	57.7	65.9	63.5	67.9
l=8	52.4	60.8	69.4	69.9	71.6	62.6	71.9
l=16	60.2	85.9	86.8	82.3	90.4	87.2	91.6

6.4.5 LFW Database

For this database, 100 subjects of LFWa database [107] are selected. 6 images are used For each person. The selected images are cropped to exclude the background, and resized to dimensions of 64×64 pixels.

Table 6.3: Recognitionrates (%) of state-of-the-art and the proposed methods on Extended YaleB face database with 20% corruption

Training samples	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
l=2	47.0	61.3	58.4	57.6	65.9	62.4	68.3
l=4	50.1	64.3	61.3	60.8	68.9	65.2	70.9
l=8	53.7	73.9	71.8	72.8	72.5	74.6	73.9
l=16	68.3	90.5	84.8	87.3	87.0	90.9	91.2

The experiments on LFW database are also carried out with random corruption and random block occlusion as shown in figures 6.7 and 6.8. Table 6.4 shows that the proposed method has the highest recognition rates among all methods with LFW database.

Table 6.4: Recognition rates (%) of state-of-the-art and the proposed methods on LFW database

Database	PCA	LFDA	LPP	SPP	CRP	SVM	Proposed method
LFW	66.9	94.2	92.7	93.8	95.2	94.6	95.6
Occluded LFW	61.1	86.9	81.4	86.2	83.2	87.5	90.1
Corrupted LFW	61.7	88.7	85.9	87.5	84.2	88.6	89.9

Figure 6.7: Samples with corruption in LFW database

Figure 6.8: Samples with occlusion in LFW database

6.5 Conclusion

A new regression based method for feature extraction and face recognition is proposed. The proposed method uses ℓ_0-norm constraint to enforce the image transformation to be sparse. ℓ_2-norm regularization terms are used to increase the similarity between the transformation of the same class, and to decrease the similarity between transformation of different class. The proposed method has been tested under various face databases. All the obtained simulation results show the superiority our method. Simulations also show the high performance of the method under image corruption and occlusion.

Chapter 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

Recognition accuracy and computational complexity are the two main factors used to measure the performance and efficiency of any face recognition system. Therefore, to develop a satisfactory method, researcher should focus on these two parameters without de-emphasizing any of them.

In this thesis, three face recognition algorithms were developed. The proposed methods were of regression based family of feature extraction and face recognition, where a dictionary is trained to transform the images into a specific sparse vector to be used in classification process. The proposed methods aim to project the face image into discriminative form, which can be considered as feature extraction method. Also, every method is integrated with a classifier that depends on ℓ_2-norm computation of the sparse representation of the image in the transform domain to discover the image class.

To evaluate the proposed algorithms, four benchmark databases were used. ORL, Extended Yale B and AR were used to evaluate the first algorithm presented in chapter 3. ORL, Extended Yale B, AR and LFWa were used to test the second algorithm explained in chapter 4. The third algorithm presented in chapter 5 was tested by ORL, Extended Yale B, AR and LFWa databases. These databases consist

of face images with various facial variations, like lightning conditions, rotation angle, facial expressions, etc.

The proposed methods are distinguished with their low computational cost and fast classification process, which only consist of one matrix-vector multiplication and norm computations. Recognition accuracies of the proposed algorithms are high and satisfactory. The experimental results that have been carried out prove that our methods have high and distinguished accuracy, especially the second and third methods. Methods two and three, which are presented in chapters 4 and 5, show robustness under image occlusion and corruption.

7.2 Future Work

The proposed methods that have been discussed in this thesis show promising results in face recognition field. We will study the ability to extend our work to more challenging conditions and with other applications than face recognition. In this section, we will discuss other ideas for the future work:

- We will modify and apply the proposed methods on different disciplines. In medical applications, for example, we believe that these algorithms are efficient methods for recognizing Alzheimer disease and different kinds of tumors.

- We will apply the proposed methods under more challenging conditions and environments such as bigger databases, low resolution images and corrupted images.

- We will study the effect of integrating the proposed methods as a pre-processing step for other methods in deep learning.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] A. K. Jain, B. Klare, and U. Park, "Face recognition: Some challenges in forensics," in *Face and Gesture 2011*, March 2011, pp. 726–733.

[3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, Dec. 2003.

[4] A. Qudaimat and H. Demirel, "Sparsifying transform learning for face image classification," *Electronics Letters*, vol. 54, no. 17, pp. 1034 – 1036, August 2018.

[5] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun 1991, pp. 586–591.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Jul 1997.

[7] X. He and P. Niyogi, "Locality preserving projections," *In:Thrun, S., Saul, K., Schölkopf, B. (Eds.), Advances in Neural Information Processing Systems*, vol. 16, pp. 153–160, 2004.

[8] J. Soldera, C. A. R. Behaine, and J. Scharcanski, "Customized orthogonal locality preserving projections with soft-margin maximization for face recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 9, pp. 2417–2426, Sept 2015.

[9] M. E. Ashalatha, M. S. Holi, and P. R. Mirajkar, "Face recognition using local features by lpp approach," in *International Conference on Circuits, Communication, Control and Computing*, Nov 2014, pp. 382–386.

[10] J. Zhang and J. Wang, "Dimensionality reduction using sparse locality preserving projections and its application in face recognition," in *2018 37th Chinese Control Conference (CCC)*, July 2018, pp. 9011–9015.

[11] Y. Li, "Locally preserving projection on symmetric positive definite matrix lie group," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 1217–1221.

[12] R. R. Singh, R. K. Khandelwal, and M. Chavan, "Face recognition using orthogonal locality preserving projections," in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, Oct 2016, pp. 1323–1328.

[13] Y. Wen, S. Yang, L. Hou, and H. Zhang, "Face recognition using locality sparsity preserving projections," in *2016 International Joint Conference on Neural Networks (IJCNN)*, July 2016, pp. 3600–3607.

[14] D. Jing and L. Bo, "Distance-weighted manifold learning in facial expression recognition," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, June 2016, pp. 1771–1775.

[15] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, ser. IJCAI'11. AAAI Press, 2011, pp. 1294–1299.

[16] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, Nov 2002.

[17] Z. Lihong, W. Ye, and T. Hongfeng, "Face recognition based on independent component analysis," in *2011 Chinese Control and Decision Conference (CCDC)*, May 2011, pp. 426–429.

[18] K. Kinage and S. Bhirud, "Face recognition using independent component analysis of gaborjet (gaborjet-ica)," 05 2010, pp. 1–6.

[19] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge, "Recognizing faces with pca and ica," *Comput. Vis. Image Underst.*, vol. 91, no. 1, pp. 115–137, Jul. 2003.

[20] P. C. Yuen and J. Lai, "Face representation using independent component analysis," *Pattern Recognition*, vol. 35, no. 6, pp. 1247–1257, 2002.

[21] X. Zhang and X. Ren, "Two dimensional principal component analysis based independent component analysis for face recognition," in *2011 International Conference on Multimedia Technology*, July 2011, pp. 934–936.

[22] J. J. Zhang and Y. T. Shi, "Face recognition systems based on independent component analysis and support vector machine," in *2014 International Conference on Audio, Language and Image Processing*, July 2014, pp. 296–300.

[23] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 156–171, Jan 2017.

[24] F. Zhang, J. Yang, J. Qian, and Y. Xu, "Nuclear norm-based 2-dpca for extracting features from images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2247–2260, Oct 2015.

[25] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, N. Sebe, and A. G. Hauptmann, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec 2012.

[26] Z. Ma, Y. Yang, N. Sebe, K. Zheng, and A. G. Hauptmann, "Multimedia event detection using a classifier-specific intermediate representation," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1628–1637, Nov 2013.

[27] Z. Lai, Y. Xu, J. Yang, L. Shen, and D. Zhang, "Rotational invariant dimensionality reduction algorithms," *IEEE Transactions on Cybernetics*, vol. 47, no. 11, pp. 3733–3746, Nov 2017.

[28] X. Chen, G. Yuan, F. Nie, and M. Zhong, "Semi-supervised feature selection via sparse rescaled linear square regression," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2018.

[29] S. Xu, J. Dai, and H. Shi, "Semi-supervised feature selection based on least square regression with redundancy minimization," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–8.

[30] W. Mao, W. Xu, and Y. Li, "Sparse feature grouping based on$\ell_{1/2}$norm regularization," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 1045–1051.

[31] T. Pang, F. Nie, J. Han, and X. Li, "Efficient feature selection via$l_{2,0}$-norm constrained sparse regression," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2018.

[32] Z. Lai, D. Mo, J. Wen, L. Shen, and W. Wong, "Generalized robust regression for jointly sparse subspace learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2018.

[33] Z. Lai, D. Mo, W. K. Wong, Y. Xu, D. Miao, and D. Zhang, "Robust discriminant regression for feature extraction," *IEEE Transactions on*

Cybernetics, vol. 48, no. 8, pp. 2472–2484, Aug 2018.

[34] R. Shang, W. Wang, R. Stolkin, and L. Jiao, "Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 793–806, Feb 2018.

[35] Z. Zhang, Z. Zhong, J. Cui, and L. Fei, "Learning robust latent subspace for discriminative regression," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, Dec 2017, pp. 1–4.

[36] Y. Lu, J. Liu, X. Kong, and J. Shang, "A convex multi-view low-rank sparse regression for feature selection and clustering," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov 2017, pp. 2183–2186.

[37] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–741, May 1995.

[38] P. D. Wadkar and M. Wankhade., "Face recognition using discrete wavelet transforms," *International Journal of Advanced Engineering Technology*, vol. 3, no. 6, pp. 239–242, 2012.

[39] A. S. B. Mahajan and K. J. Karande, "Pca and dwt based multimodal biometric recognition system," in *2015 International Conference on Pervasive Computing (ICPC)*, Jan 2015, pp. 1–4.

[40] M. M. Mukhedkar and S. B. Powalkar, “Fast face recognition based on wavelet transform on pca,” in *2015 International Conference on Energy Systems and Applications*, Oct 2015, pp. 761–764.

[41] M. Luo, L. Song, and S. Li, “An improved face recognition based on ica and wt,” in *2012 IEEE Asia-Pacific Services Computing Conference*, Dec 2012, pp. 315–318.

[42] K. S. Kinage and S. G. Bhirud, “Face recognition based on independent component analysis on wavelet subband,” in *2010 3rd International Conference on Computer Science and Information Technology*, vol. 9, July 2010, pp. 436–440.

[43] X. Zhihua and L. Guodong, “Weighted infrared face recognition in multiwavelet domain,” in *2013 IEEE International Conference on Imaging Systems and Techniques (IST)*, Oct 2013, pp. 70–74.

[44] A. Aldhahab, G. Atia, and W. B. Mikhael, “Supervised facial recognition based on multi-resolution analysis and feature alignment,” in *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug 2014, pp. 137–140.

[45] A. Aldhahab, G. Atia, and W. Mikhael, “Supervised facial recognition based on eigenanalysis of multiresolution and independent features,” in *2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug 2015, pp. 1–4.

[46] A. A. G. Azzawi and M. A. H. Al-Saedi, “Face recognition based on mixed between selected feature by multiwavelet and particle swarm optimization,” in *2010 Developments in E-systems Engineering*, Sept 2010, pp. 199–204.

[47] A. Aldhahab, G. Atia, and W. B. Mikhael, “Supervised facial recognition based on multiresolution analysis with radon transform,” in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 928–932.

[48] Z. Karhan and B. Ergen, “Classification of face images using discrete cosine transform,” in *2013 21st Signal Processing and Communications Applications Conference (SIU)*, April 2013, pp. 1–4.

[49] D. Sisodia, L. Singh, and S. Sisodia, “Incremental learning algorithm for face recognition using dct,” in *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, March 2013, pp. 282–286.

[50] S. Ajitha, A. A. Fathima, V. Vaidehi, M. Hemalatha, and R. Karthigaiveni, “Face recognition system using combined gabor wavelet and dct approach,” in *2014 International Conference on Recent Trends in Information Technology*, April 2014, pp. 1–6.

[51] F. Z. Chelali, A. Djeradi, and N. Cherabit, “Investigation of dct/pca combined with kohonen classifier for human identification,” in *2015 4th International Conference on Electrical Engineering (ICEE)*, Dec 2015, pp. 1–7.

[52] A. Dahmouni, N. Aharrane, K. E. Moutaouakil, and K. Satori, "Face recognition using local binary probabilistic pattern (lbpp) and 2d-dct frequency decomposition," in *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*, March 2016, pp. 73–77.

[53] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov 2015.

[54] Z. Lei, D. Yi, and S. Z. Li, "Learning stacked image descriptor for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1685–1696, Sept 2016.

[55] W. Ouarda, H. Trichili, A. M. Alimi, and B. Solaiman, "Mlp neural network for face recognition based on gabor features and dimensionality reduction techniques," in *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, April 2014, pp. 127–134.

[56] Z. Zhang, J. Li, and R. Zhu, "Deep neural network for face recognition based on sparse autoencoder," in *2015 8th International Congress on Image and Signal Processing (CISP)*, Oct 2015, pp. 594–598.

[57] V. E. Liong, J. Lu, and G. Wang, "Face recognition using deep pca," in *2013 9th International Conference on Information, Communications Signal Processing*, Dec 2013, pp. 1–5.

[58] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb 2009.

[59] J. Yang, L. Zhang, Y. Xu, and J. Y. Yang, "Beyond sparsity: The role of $\ell_1 - optimizer$ in pattern classification," *Pattern Recognition*, vol. 45, no. 3, pp. 1104–1118, 2012.

[60] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *2011 International Conference on Computer Vision*, Nov 2011, pp. 471–478.

[61] L. Qiao, S. Chen, , and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, 2010.

[62] W. Yang, Z. Wang, and C. Sun, "A collaborative representation based projections method for feature extraction," *Pattern Recogition*, vol. 48, no. 1, pp. 20–27, 2015.

[63] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *CVPR 2011*, June 2011, pp. 625–632.

[64] Q. Gao, Q. Wang, Y. Huang, X. Gao, X. Hong, and H. Zhang, "Dimensionality reduction by integrating sparse representation and fisher criterion and its applications," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp.

5684–5695, Dec 2015.

[65] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals Eugen. 7*, pp. 179–188, 1936.

[66] C.-Y. Lu, H. Min, J. Gui, L. Zhu, and Y.-K. Lei, "Face recognition via weighted sparse representation," *Journal of Visual Communication and Image Representation*, vol. 24, no. 2, pp. 111 – 116, 2013, sparse Representations for Image and Video Analysis.

[67] X. Song, Z. Liu, X. Yang, and S. Gao, "A new sparse representation-based classification algorithm using iterative class elimination," *Neural Computing and Applications*, vol. 24, no. 7, pp. 1627–1637, 2014.

[68] L. He, H. Li, Q. Zhang, and Z. Sun, "Dynamic feature matching for partial face recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 791–802, Feb 2019.

[69] S. Yang, L. Zhang, L. He, and Y. Wen, "Sparse low-rank component-based representation for face recognition with low-quality images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 1, pp. 251–261, Jan 2019.

[70] Y. Wang, Y. Y. Tang, L. Li, H. Chen, and J. Pan, "Atomic representation-based classification: Theory, algorithm, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 6–19, Jan 2019.

[71] R. Serajeh and A. Mousavinia, "Single sample face recognition: Discriminant scaled space vs sparse representation-based classification," in *2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)*, Oct 2018, pp. 286–292.

[72] M. Melek, A. Khattab, and M. F. Abu-Elyazeed, "Fast matching pursuit for sparse representation-based face recognition," *IET Image Processing*, vol. 12, no. 10, pp. 1807–1814, 2018.

[73] J. Mounsef and L. Karam, "Augmented sparse representation classifier for blurred face recognition," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 778–782.

[74] J. Gao and L. Zhang, "Improved face recognition based on the fusion of pca feature extraction and sparse representation," in *2018 Chinese Control And Decision Conference (CCDC)*, June 2018, pp. 3473–3479.

[75] W. Deng, J. Hu, and J. Guo, "Face recognition via collaborative representation: Its discriminant nature and superposed representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2513–2521, Oct 2018.

[76] M. Wu, S. Li, and J. Hu, "Extended class-wise sparse representation for face recognition," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Dec 2017, pp. 1611–1615.

[77] S. Yang and Y. Wen, "A novel src based method for face recognition with low quality images," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 3805–3809.

[78] M. Ahmed, M. Albashier, A. Tan, A. Abdaziz, and F. Sammani, "An effective method to tackle illumination problem in collaborative representation based classification," in *TENCON 2017 - 2017 IEEE Region 10 Conference*, Nov 2017, pp. 2209–2213.

[79] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell_1 - minimization$," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, mar 2003.

[80] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal $\ell_1 - norm$ solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[81] E. Candes, J. Romberg, and T. Tao, "Stable Signal Recovery from Incomplete and Inaccurate Measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[82] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, Dec 2006.

[83] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20,

no. 3, pp. 273–297, Sep 1995.

[84] V. Mehta, S. Khandelwal, and A. K. Kumawat, "A survey on face recognition algorithm," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, May 2018, pp. 1–3.

[85] H. Jia and A. M. Martinez, "Support vector machines in face recognition with occlusions," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 136–141.

[86] N. Lopes, A. Silva, S. R. Khanal, A. Reis, J. Barroso, V. Filipe, and J. Sampaio, "Facial emotion recognition in the elderly using a svm classifier," in *2018 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)*, June 2018, pp. 1–5.

[87] M. Jiu, N. Pustelnik, and L. Qi, "Multiclass svm with hierarchical interaction: Application to face classification," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2018, pp. 1–6.

[88] M. K. Bhuyan, S. Dhawle, P. Sasmal, and G. Koukiou, "Intoxicated person identification using thermal infrared images and gait," in *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, March 2018, pp. 1–3.

[89] L. O. Jimenez and D. A. Landgrebe, "Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of

multivariate data," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 28, no. 1, pp. 39–54, Feb 1998.

[90] P. W. Hallinan, "A low-dimensional representation of human faces for arbitrary lighting conditions," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Jun 1994, pp. 995–999.

[91] H. Murase and S. K. Nayar, "Visual learning and recognition of 3-d objects from appearance," *International Journal of Computer Vision*, vol. 14, no. 1, pp. 5–24, 1995.

[92] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, pp. 1–30, Jun 2006.

[93] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, Dec 1994, pp. 138–142.

[94] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, Jun 2001.

[95] A. Martinez and R. Benavente, "The ar face database," Computer Vision Center, Tech. Rep. 24, June 1998.

[96] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[97] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, May 2005.

[98] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1072–1086, March 2013.

[99] R. Rubinstein, T. Faktor, and M. Elad, "K-svd dictionary-learning for the analysis sparse model," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 5405–5408.

[100] M. Yaghoobi, S. Nam, R. Gribonval, and M. E. Davies, "Noise aware analysis operator learning for approximately cosparse signals," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 5409–5412.

[101] M. Yaghoobi, T. Blumensath, and M. E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Transactions on Signal Processing*, vol. 57, no. 6, pp. 2178–2191, June 2009.

[102] M. Aharon, M. Elad, and A. Bruckstein, "lt;tex gt;*rmk* lt;/tex gt;-svd: An algorithm for designing overcomplete dictionaries for sparse representation,"

IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[103] R. Pytlak, *Conjugate Gradient Algorithms in Nonconvex Optimization*. Springer, 2009.

[104] J. Dattorro, *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing USA, 2005.

[105] E. Candes and J. Romberg, "ℓ_1-magic: Recovery of sparse signals via convex programming," https://statweb.stanford.edu/candes/l1magic/, 2005 (accessed September 5, 2016).

[106] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recogn.*, vol. 43, no. 1, pp. 331–341, Jan. 2010.

[107] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 3730–3738.