Sparse representation based classification for face recognition

Lai, Jian

2015

Lai, J. (2015). Sparse representation based classification for face recognition. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/65460>

<https://doi.org/10.32657/10356/65460>

Sparse Representation Based Classification for Face Recognition

LAI Jian

School of Electrical & Electronic Engineering

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

2015

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

.....

Date

.....

LAI Jian

Acknowledgements

Foremost, I would like to express my heartfelt gratitude to my advisor, Associate Professor JIANG Xudong, for his professional guidance, unconditional support and continuous encouragement through my four years of Ph.D. study and research. It has been an honor for me to be one of his Ph.D. students. His immense knowledge, inspiring nature, enthusiastic assistance, scientific vision, high criterion, and detailed comments have been great sources of motivation throughout my doctoral studies. Without his numerous help, suggestions and discussions, I would not have completed this research work. His sharp insights, criticism and strict attitude towards the research work have been a resource and model for me to complete this thesis and they will continuously benefit me throughout my lifetime.

I would like to thank Dr. MIAO Zhenwei, Dr. GENG Cong , Dr. REN Jianfeng, Dr. AMIT Satpathy and the students and staffs in Workstation Resource Lab for their help. I am grateful to Mr. CHU Stewart and Mr. YANG Yongseng for their kind technical supports.

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

Finally, I must thank my family. I would like to dedicate this work to my parents and wife for their persistent love, continuous encouragement, endless support.

Abstract

Building a computer as intelligent as or more intelligent than human is the ultimate goal of machine learning. To explore this powerful ability in computer vision, it inspires numerous interesting and useful research areas and applications. Face recognition (FR), as one of the most challenging problems in machine learning and computer vision, provides a good test bed to show how intelligent a machine can be. Moreover, it shows huge potential for commercial and surveillance, such as access control, computer-human interaction and anti-terrorist. FR remains a hot research topic though it has been studied for more than three decades. One important reason is that there is still large gap between human and machine to recognize the identity of a face image due to extreme illumination, expression and pose variation, disguise, and occlusion *et al.*.

Recently, sparse representation based classification (SRC) has been introduced and applied to FR in dealing with some challenge scenarios. Different from the traditional classifiers, which represent the query image by each individual class, SRC considers the query image as a sparse linear combination of training samples of all subjects. The optimization of the sparse coefficients lets every training sample compete against the others to win its representation share of the query image. This is a good discriminating process. The advantages of SRC lead to some encouraging and impressive FR results. Although SRC shows its merit in handling the query image with sparse corruption (e.g., occlusion or disguise), there are still many problems remained to be addressed. In this thesis, we propose several algorithms for FR to tackle the limitations

of SRC. Comprehensive results to validate our proposed feature extraction and classification frameworks are given.

Very recently, SRC has been applied to feature extraction or dimensionality reduction. We first analyze the limitation of unsupervised procedure of SRC in dimensionality reduction. To address this problem, we propose a supervised sparse representation (SSR) utilizing training label information. By promoting the coefficients of correct class and suppressing those of the others, the proposed SSR learns a more discriminative representation than SRC does. Furthermore, we also propose a dimensionality reduction method, named discriminative sparsity preserving embedding (DSPE). The proposed DSPE captures the discriminative sparse and discriminative structure of labelled training data based on the proposed SSR and finds the low dimensional subspace that reduces the within class distances meanwhile keeping the between class distances.

We then develop a modular weighted global sparse representation (WGSR) framework to handle the severely corrupted query image. In the proposed WGSR, an image is first divided into modules and each module is processed separately to determine its reliability. We propose to use the modular sparsity and residual jointly to determine the modular reliability. A reconstructed image from the modules weighted by their reliability is formed for the robust recognition. The proposed framework advances both the modular and global sparse representation approaches, especially, in dealing with query images with disguise, large illumination variations or expression changes.

The third contribution of this thesis is the introduction of a class-wise sparse representation with collaborative patch (CSR-CP). To alleviate the problem of unsupervised optimization of sparse representation, we proposed a class-wise sparse representation (CSR) by employing class label information. The proposed CSR seeks an optimum representation of the query image such that it simultaneously minimizes the reconstruction error and maximizes class-wise sparsity. To tackle the problem of insufficient training data, we further propose a collaborative patch

framework, together with CSR, named CSR-CP. Different from the patch based methods, which optimize each patch independently, the proposed CSR-CP reconstructs the whole query sample by sub-images of all subjects across all pathes within one framework, and searches a patch-wise and class-wise sparse representation. As a result, a more reliable and discriminative result can be attained.

Finally, we addresses two fundamental problems of SRC by proposing a sparse- and dense-hybrid representation (SDR) via the supervised low rank (SLR) dictionary decomposition approach. With sufficient representative and well control training data, SRC has proven its success in some challenging scenarios. A violation of these two conditions leads to a poor performance of SRC. However, FR is an application where we have a large number of subjects but sufficient representative and uncorrupted training images cannot be guaranteed for every subject. To alleviate the problems of SRC, a sparse- and dense-hybrid representation (SDR) framework is proposed. We further propose a procedure of supervised low-rank (SLR) dictionary decomposition to facilitate the proposed SDR framework. In addition, the problem of the corrupted training data is also alleviated by the proposed SLR dictionary decomposition. The application of the proposed SDR-SLR approach in face recognition verifies its effectiveness and advancement to the field.

Contents

Acknowledgements	i
Abstract	ii
List of Figures	xi
List of Tables	xvi
1 Introduction	1
1.1 Motivation	1
1.2 Objectives	5
1.3 Contributions	6
1.4 Dissertation Organization	8
2 Literature Review	9
2.1 Classifiers in Face Recognition	9
2.1.1 Nearest Neighbor	10
2.1.2 Nearest Feature Line	11
2.1.3 Nearest Feature Plane	11

2.1.4	Nearest Feature Subspace	13
2.1.5	Linear Regression Classification	13
2.1.6	Sparse Representation based Classification	16
2.1.7	Extension of SRC	19
2.2	Linear Feature Extraction in Face Recognition	23
2.2.1	Principle Component Analysis	24
2.2.2	Linear Discriminative Analysis	24
2.2.3	Neighborhood Preserving Embedding	25
2.2.4	Sparsity Preserving Projections	26
2.3	Summary	27
3	Discriminative Sparsity Preserving Embedding for Face Recognition	29
3.1	Introduction	29
3.2	Proposed Framework for Feature Extraction	31
3.2.1	Supervised Sparse Representation	32
3.2.2	Preserving the Supervised Sparse Reconstructive Coefficients	36
3.3	Experiments	39
3.3.1	Yale Face Database	40
3.3.2	ORL Face Database	41
3.3.3	AR Database	41
3.3.4	CMU PIE Database	42
3.3.5	Effect of θ	43

3.4	Summary	45
4	Modular Weighted Global Sparse Representation for Robust Face Recognition	47
4.1	Introduction	48
4.2	Proposed Modular Weighted Global Sparse Representation	50
4.2.1	Determination of the Modular Weights	50
4.2.2	Weighted Global Sparse Representation	54
4.2.3	Classification Procedure	56
4.3	Experiments	57
4.3.1	Extended Yale B Database	58
4.3.2	AR Database	59
4.3.3	Multi-PIE	61
4.4	Summary	63
5	Class-wise Sparse and Collaborative Patch Representation for Robust Face Recognition	65
5.1	Introduction	66
5.2	Class-wise Sparse Representation	68
5.2.1	Optimization	74
5.2.2	Classification	77
5.3	Collaborative Patch Representation	78
5.4	Experiments	84
5.4.1	Face Recognition with Uncorrupted Training Samples	86

5.4.2	Face Recognition with Occluded Training Samples	89
5.4.3	Face Recognition with Different Patch Numbers	92
5.5	Summary	93
6	Sparse and Dense Hybrid Representation via Dictionary Decomposition for Face Recognition	95
6.1	Introduction	96
6.2	Sparse and Dense Hybrid Representation	99
6.3	Dictionary Decomposition	107
6.4	Experiments	117
6.4.1	Face Recognition on Single Variation	119
6.4.2	Face Recognition on Mixed Variations	124
6.4.3	Face Recognition on Corrupted Data	126
6.4.4	Face Recognition with Real Disguised Images	128
6.5	Summary	131
7	Conclusions and Future Work	133
7.1	Contributions	134
7.2	Future Work	136
	Author's Publications	138
	Bibliography	140

List of Figures

1.1	A generic face recognition system.	2
1.2	Some examples of the biometric characteristics:(a) iris; (b) fingerprint; (c) palm-print; (d) ear; (e) gait.	3
2.1	The feature line generalized by two prototype feature samples $\mathbf{a}_{i,j}$ and $\mathbf{a}_{i,k}$. The query sample $\mathbf{y}$ is projected onto the feature line.	11
2.2	Feature plane consisting of $\mathbf{a}_{i,l}$, $\mathbf{a}_{i,k}$, $\mathbf{a}_{i,g}$. The query sample $\mathbf{y}$ is projected onto the plane.	12
2.3	Partitioned face image with 4, 6, and 8 modules.	14
2.4	Overview of SRC approach. SRC represents a query image $\mathbf{y}$ (a), which a partially occluded, as a sparse linear combination $\mathbf{x}_1$ (b) of all the training sample $\mathbf{A}$ (c) plus a sparse error $\mathbf{e}_1$ (d). The red coefficients correspond to training images of the correct individual.	20
3.1	pdf of Laplace distribution with different selections of λ	34
3.2	Distribution of nonzero coefficients in weighting matrices: (a) matrix of NPE, (b) matrix of SPP, (c) matrix of DSPE.	37
3.3	Recognition rates on AR database versus number of dimensionality.	42

3.4	Recognition rates on CMU PIE database. The results of 10, 20 and 30 training samples per subject are shown in (a), (b) and (c), respectively.	44
3.5	Recognition rates versus θ on CMU PIE database.	45
4.1	The first row shows normal images. The second row shows image with variations caused by disguise, expression and large illumination changes.	49
4.2	(a) A query face image with scarf. The first row of (b)-(i) shows the eight modules of query image. The second row of (b)-(i) shows the reconstructed image with ℓ_1 -norm minimization. The third row of (b)-(i) shows the estimated modular class coefficients, and the numeric values of modular sparsity and residual. The red coefficients (circle) correspond to correct training subject.	51
4.3	(a) Weighting function of modular sparsity; (b) Weighting function of modular residual.	55
4.4	Recognition rate on Extended Yale B versus number of training sample per subject.	60
4.5	The query images and their corresponding weight images and weighted global query images. (a) Query image with scarf; (b) weight image of (a); (c) the weighted global query image of (a); (d) query image with sunglasses; (e) weight image of (d); (f) the weighted global query image of (d). In (b) and (e), black and white pixels denote 0 and 1 weight, respectively.	62
4.6	Faces from Multi-PIE database with different expressions.	63
5.1	An example of two different representations: a query image y (circle) can be sparsely well represented by samples $a_{1,1}$ and $a_{1,2}$ of one class (triangle); it can also be densely well represented by samples $a_{2,1}$ and $a_{2,2}$ of another class (diamond).	71

5.2	Comparison of representation coefficients of SRC (a), GL (b), and the proposed CSR (c) on AR database of 20 persons with 5 undisguised training images per person. The x-axis is $5(k-1) + j$ where k and j are respectively the indices of subjects and training images of a subject. Coefficients of the correct subject are in red lines (circles) and others are in blue lines (dots).	73
5.3	(a) A query image partitioned into the predefined patches. (b) An example of collaborative patch sparse representation of the whole query image. The coefficients indicated by red lines (circle) are associated with the correct class. .	79
5.4	Comparison of SRC (a), SRC-P (b) and the proposed CSR-CP (c) in terms of representation accuracy by single class. For SRC-P, the class residual r_i is the sum of the 8 patch residuals for comparison with SRC and CSR-CP. The red lines (circle) are associated with the correct subject.	82
5.5	Face samples from the Extended Yale B.	85
5.6	Face samples from the AR database.	85
5.7	Face samples from the CMU PIE database.	86
5.8	Face samples with illumination variation from CMU Multi-PIE database. . . .	87
5.9	Recognition rate on Multi PIE Database.	88
5.10	The convergence curve of $\ \mathbf{y} - \mathbf{Ax} - \mathbf{e}\ _\infty$ versus iteration number.	91
5.11	Face samples partitioned into different numbers of patch. (a) 2×2 ; (b) 3×2 ; (c) 4×2 ; (d) 5×2 (e) 6×2	92
5.12	Recognition rate on AR database with different numbers of patch.	93
6.1	Image decomposition: a face image projected to the class-specific subspace, the non-class-specific subspace and their residual random sparse noise.	103

6.2	Comparison of SRC (1th row) and the proposed SDR (2nd row) on AR database A of 100 persons with 3 images per person: (a) query image; (b) SRC coefficients of A and SDR coefficients of D and B ; (c) top 16 images in A of SRC and in D and B of SDR; (d) residuals. Sparse coefficients of the correct person are in red lines and its training samples are in yellow rectangles. Images are normalized into the same mean and ℓ_2 -norm.	104
6.3	An example of SLR's convergence on the AR database.	115
6.4	Sample images produced by the proposed SLR: images of a person in A corrupted by random noise (a); their initial components assigned to D (b) and B (c); final images in D (d), B (e) and E (f).	115
6.5	Examples of RPCA and the proposed SLR: training images (a); low-rank (b) and sparse (c) images from RPCA; class-specific (d), non-class-specific (e) and sparse (f) images decomposed by SLR.	116
6.6	Normalized images from FERET database.	118
6.7	Sample images in Multi-PIE database with variations of (a) illumination, (b) expression, (c) pose.	120
6.8	Recognition rate versus number of training illuminations per subject on Multi-PIE database.	121
6.9	Recognition rate versus number of training expressions per subject on Multi-PIE database.	122
6.10	Recognition rate versus percentage of uniform noise on AR database.	127
6.11	Examples of images of Extended Yale B database with different levels of block occlusion. From left to right, 20%, 30%, 40% and 50% of images are occluded, respectively.	128

6.12 Recognition rate versus percentage of occlusion on extended Yale B database. .	129
---	-----

List of Tables

3.1	Recognition Rate on the Yale Face Database.	40
3.2	Recognition Rate on the ORL Face Database.	41
4.1	Recognition Rate on the Extended Yale B Database.	58
4.2	Recognition Rate on the AR Database.	61
4.3	Recognition Rate on the Multi-PIE Database.	63
5.1	Recognition Rate on Uncorrupted Extended Yale B.	87
5.2	Recognition Rate on AR Face Database.	89
5.3	Recognition Rate on CMU PIE Database.	90
5.4	Recognition rate on AR database with only disguised training samples.	90
5.5	Recognition Rate on Disguised AR Face Database.	92
6.1	Recognition Rate on Extended Yale B	120
6.2	Recognition Rate on the Multi-PIE Face Database with Pose Variation.	123
6.3	Recognition Rate on All Undisguised Images of AR Dataset	124
6.4	Recognition Rate of Fewer Training Samples of AR Dataset	125

6.5	Recognition Rate on FERET Database	126
6.6	Recognition Rate on the AR Database.	131

Chapter 1

Introduction

1.1 Motivation

A 2-D face recognition system is a computer application of learning the identity of a subject based on a still image or video frame. In general, it consists of 4 modules: face detection, face alignment, feature extraction and classification. Fig. 1.1 illustrates the whole process of face recognition. Typically, given a image, the facial area is cropped out based on the detected location and scale of face. The cropped face is aligned according to the facial landmarks (e.g., eyes, nose, and mouth), which is followed by the feature extraction. In feature extraction, much lower dimensionality features are extracted from the high dimensional aligned faces for both computational and accuracy concern. Finally, the input face is classified to one of those enrolled users in database.

In addition to face [1–18], biometric recognition approaches [19–21] also recognize a person based on other physiological or behavioral characteristics, such as iris [22–25], fingerprint [26–30], palmprint [31, 32], ear [33, 34] and gait [35–37] shown in Fig. 1.2. Compared with other biometric approaches, face recognition has the several advantages. First, it requires only a digital camera for collection of face data. However, to capture the iris, fingerprint or palmprint,

Figure 1.1: A generic face recognition system.

specific equipments are needed. Further more, face recognition can be carried out without the perception of the subject, which is an user friendly system. Nevertheless, collection of fingerprint requires the physical contact with the sensor and that of iris needs the subject to stand just before the device, which are not feasible in many realistic scenarios, especially in surveillance applications. Moreover, face recognition can perform in the mass environment, such as airport and train station. Faces can be recorded in a long distance and many faces can be captured in one image, where it may be not possible for the others.

As a result of these advantages, face recognition has huge commercial and surveillance potential. From the aspects of safety and convenience, door entry systems could utilize face recognition to replace the traditional personal identification number (PIN), which is easily peeped. Face-based access control has already been employed in PC and mobile phone. Face recognition is also applied to human-computer interaction for video games, such as XBOX Kinect.

Figure 1.2: Some examples of the biometric characteristics:(a) iris; (b) fingerprint; (c) palm-print; (d) ear; (e) gait.

In addition, thanks to the existing vast surveillance cameras in the public places, the recorded images and videos could be directly fed into face recognition system to locate the person in the watchlist. What is more, as there is an increasing popularity of image and video sharing social network, such as Facebook and Youtube, it will be useful to apply the face recognition to organize the large amount of visual data.

As one of the most challenging problems in machine learning and computer vision, face recognition has attracted considerable attention of the numerous researchers for more than 30 years. Although it is a trivial task for human, identifying a person by a face image is still an extremely difficult task for machine due to the following reasons:

- Large variation of illumination, pose, and expression. These variations enlarge the intra-class distances, meanwhile reduces the inter-class distances. The similarity of face images from the same person with different variations could be much lower than those of the

different persons but with similar variations, where it is non-trivial to classify such query sample correctly.

- Disguise and occlusion. Face could be easily disguised with scarf, sunglasses *et. al.*. In a surveillance scenario, such as airport, face could be partially occluded by other persons or objects. In both cases, the appearance of the captured face can be significantly different from the normal sample of same subject.
- Huge number of subjects. It is a big challenge to store the face data of a large subject set. In addition, it requires a huge burden of computational power to handle such mass data. Further more, the appearances of some subjects can be very similar, which leads to less reliable results.
- Lack of sufficient representative training data. This is also called small-sample-size problem. With sufficient training data, even a simple classifier could achieve high accuracies. However, in the practical application, it is usually only a few or even one training sample per subject available, such as the ID photo, where it is time and manpower consuming to collect more data. In this case, the performances of many classic approaches will degrade dramatically.

The existing face recognition technologies can achieve high recognition rates under well controlled environment. However, they are vulnerable in uncontrolled scenarios, such as the situation discussed above. These challenging problems have driven and motivated researchers to develop various approaches.

1.2 Objectives

Inspired by wavelet, ridgelet, curvelet and contourlet transforms, sparse representation of signals has been proposed and attracted considerable attention. However, it is difficult to directly find such a sparse solution over an overcomplete dictionary via ℓ_0 -norm minimization, which is NP-hard. Fortunately, thanks to the studies of compressive sampling [38–42], the ℓ_0 -norm minimization problem can be efficiently solved by convex optimization (e.g., ℓ_1 -norm minimization [43, 44]) and greedy pursuit [41, 45–49], if the solution is sufficiently sparse.

Spare representation has shown its outstanding power for acquiring, representing and compressing. In the past a few years, it has been widely applied to fields of computer vision, which includes image super-resolution [50–52], data clustering [53–55], deblurring [56, 57], denoising [58–60], visual tracking [61–64] and background modeling [65, 66].

Recently, sparse representation based classification (SRC) [67] has been introduced to face recognition in dealing with some challenging scenarios. Different from the traditional classifiers, which evaluate the relation between the query image and the training samples of each individual class separately, SRC considers the query image as a sparse linear combination of training samples of all subjects. The optimization of the sparse coefficients lets every training sample compete against the others to win its representation share of the query image. Moreover, the ℓ_1 -norm minimization of representation error enables SRC to recognize query image sparsely corrupted by outlier pixels or occlusion. As a result, SRC achieves some impressive and interesting results.

Although sparse representation based approaches shows its merits in face recognition, there are still some problems needed to be addressed.

As sparse representation is an unsupervised procedure, sparse representation, though learns

some discriminative information of data, is not the optimum choice to capture the discriminative structure of the training data for feature extraction. Is it possible to make use of the label information to learn a more discriminative underlying structure of the training data?

SRC is robust to handle the query image with sparse corruption. However, as the portion of corruption increases, the sparse assumption of the noise or occlusion is violated that leads to a poor performances of SRC. Is it possible to learn a robust framework to deal with the severely corrupted query image?

To achieve a high accuracy, SRC requires sufficient training data. In many practical scenarios, there are large number of identities, but enough representative samples are not available for all subjects, which limits the applications of SRC. Is it possible to learn a sparse representation based framework that cooperates with insufficient training data?

One important requirement of SRC is the carefully controlled training images, which limits SRC based approaches from corrupted training data. In some realistic scenarios, the training data is corrupted during the data collection, how to utilize these training data is another problem we want to tackle.

To address the aforementioned limitations of SRC, in this thesis, we propose several sparse representation based approaches.

1.3 Contributions

The major contributions of this thesis are summarized as follows.

- A supervised sparse representation (SSR) algorithm is proposed. By utilizing the label information, SSR promotes the coefficients associated to the correct subject and suppresses

those of the others. As a result, a more discriminative representation can be learnt for each training sample. Furthermore, we propose a dimension reduction method, named discriminative sparsity preserving embedding (DSPE). The proposed DSPE captures the discriminative sparse structure with SSR, and finds the low dimensional subspace that reduces the within class distances and keeps the between class distances, which results in a more efficient embedding in terms of recognition rates.

- We develop a modular weighted global sparse representation (WGSR) framework to deal with severely occluded query face. In the proposed WGSR, an image is first divided into modules and each module is processed separately to determine its reliability. A reconstructed image from the modules weighted by their reliability is formed for the robust recognition. We propose to use the modular sparsity and residual jointly to determine the modular reliability. The proposed framework advances both the modular and global sparse representation approaches, especially in dealing with the query image with disguise, large illumination variations and expression changes.
- A class-wise sparse representation with collaborative patch (CSR-CP) approach is proposed. To alleviate the problem of unsupervised optimization of SRC, we proposed a class-wise sparsity representation (CSR) by employing class label information. The proposed CSR seeks an optimum representation of the query image such that it minimizes the reconstruction error and maximizes class-wise sparsity simultaneously. To tackle the problem of insufficient training data, we further propose a collaborative patch framework, together with CSR, named CSR-CP. Different from the patch based methods, which optimize each patch independently, the proposed CSR-CP reconstructs the whole query sample by sub-images of all subjects across all patches within one framework and searches a patch-wise and class-wise sparse representation. As a result, it alleviates the problem of losing discriminative information in the training data caused by the partition of the image into patches.

- By investigating the limitations of SRC, we propose a sparse- and dense-hybrid representation (SDR) framework with supervised low-rank (SLR) dictionary decomposition. To represent a query image, every training sample only uses its class-specific component to compete against the others (through a sparse minimization) collaboratively with the non-class-specific component of all training samples (through a dense representation). We further propose a procedure of SLR dictionary decomposition to facilitate the proposed SDR framework. In addition, the problem of the corrupted training data is also alleviated by the proposed SLR dictionary decomposition.

1.4 Dissertation Organization

The remainder of this thesis is organized as follows.

1. In Chapter 2, the related representative approaches of face recognition are reviewed.
2. In Chapter 3, supervised sparse representation (SSR) algorithm and discriminative sparsity preserving embedding (DSPE) are proposed for feature extraction.
3. In Chapter 4, a modular weighted global sparse representation (WGSR) framework is proposed.
4. In Chapter 5, a class-wise sparse representation with collaborative patch (CSR-CP) approach is proposed.
5. In Chapter 6, a sparse- and dense-hybrid representation (SDR) framework with supervised low-rank (SLR) dictionary decomposition is proposed.
6. In Chapter 7, we summarize this thesis by highlighting the contributions of our work, which is followed by the potential future work.

Chapter 2

Literature Review

In this chapter, some widely used classification and feature extraction technologies are reviewed and analyzed.

Before going through the detailed review, some notations are given. We assume that there are c subjects and each subject has n_i training samples, for $1 \leq i \leq c$. The feature vector of the j th training sample of the i th subject is denoted as $\mathbf{a}_{i,j} \in \mathbb{R}^m$, where m is the dimensionality of feature. By stacking the n_i feature vectors of the i th subject, it constructs $\mathbf{A}_i = [\mathbf{a}_{i,1}, \mathbf{a}_{i,2}, \dots, \mathbf{a}_{i,n_i}]$. $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_c] = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}$ is generated by combining the training samples of all subjects, where $n = \sum_{i=1}^c n_i$. The query sample is denoted by $\mathbf{y} \in \mathbb{R}^m$.

2.1 Classifiers in Face Recognition

Although there are many different classifiers available, the suitable ones for high dimensional face image data are limited. Nearest neighbor (NN) classifier is widely applied for its simplicity. However, NN uses the training samples as isolated points, which may not well capture the variations within class. To alleviate this problem, various classifiers are proposed to extend the

feature points to feature lines, planes and subspaces. The commonalities of all these classifiers are that they are not robust to the image corruption caused by the outlier, and that they evaluate the relation between the query image and the training samples of each individual class one by one separately. Different from the aforementioned classifiers, sparse representation based classification (SRC) considers the query image as a sparse linear combination of training samples of all classes. It lets the training sample compete with the others to win its share in the representation of query image. This is a good discriminating process. Moreover, the ℓ_1 -norm minimization of the representation error enables SRC to recognize query images corrupted by sparse outlier and occlusions. Therefore, SRC achieves some impressive results, which attract great interest for further extensions.

We give the literature review of these classifiers in the following subsections.

2.1.1 Nearest Neighbor

The nearest neighbor (NN) is a simple and popular classifier in many research fields. NN measures the Euclidean distances between the query image and each training sample of class i , and denotes the minimum distance of the i th class as $r_i(\mathbf{y})$, which can be formulated as

$$r_i(\mathbf{y}) = \min_j \|\mathbf{y} - \mathbf{a}_{i,j}\|_2. \quad (2.1)$$

The query image $\mathbf{y}$ is labelled to the class with minimum distance

$$i^* = \arg \min_i r_i(\mathbf{y}). \quad (2.2)$$

If the number of training data is limited, the isolated training samples can not well represent the whole face subspace, which degrades the performances of NN.

Figure 2.1: The feature line generalized by two prototype feature samples $\mathbf{a}_{i,j}$ and $\mathbf{a}_{i,k}$. The query sample $\mathbf{y}$ is projected onto the feature line.

2.1.2 Nearest Feature Line

To extend the isolated training samples, nearest feature line (NFL) [68] links two training samples from the same subject to construct a line shown in Fig. 2.1. The feature line goes through the feature point $\mathbf{a}_{i,j}$ and $\mathbf{a}_{i,k}$ could be formulated as

$$\mathbf{p}_i^{jk}(\alpha) = \alpha \mathbf{a}_{i,j} + (1 - \alpha) \mathbf{a}_{i,k}, \quad (2.3)$$

where α is a scalar. The optimum distance of class i is denoted as

$$r_i(\mathbf{y}) = \min_{j,k,\alpha} \|\mathbf{y} - \mathbf{p}_i^{jk}(\alpha)\|_2. \quad (2.4)$$

The same as (2.2), the query sample is classified to the class with minimum distance.

2.1.3 Nearest Feature Plane

To further extend the face subspace, nearest feature plane (NFP) [69] classifier is proposed.

Figure 2.2: Feature plane consisting of $\mathbf{a}_{i,l}$, $\mathbf{a}_{i,k}$, $\mathbf{a}_{i,g}$. The query sample $\mathbf{y}$ is projected onto the plane.

NFP introduces a plane constructed by arbitrary three samples of the same class as $\mathbf{F}_i^{lkg} = sp(\mathbf{a}_{i,l}, \mathbf{a}_{i,k}, \mathbf{a}_{i,g})$. A feature plane is shown in Fig. 2.2. A $m \times 3$ matrix is generated as $\mathbf{B}_i^{lkg} = [\mathbf{a}_{i,l}, \mathbf{a}_{i,k}, \mathbf{a}_{i,g}]$. The projection point $\mathbf{p}_i^{lkg}$ of $\mathbf{y}$ on the feature plane can be derived as

$$\mathbf{p}_i^{lkg} = \mathbf{B}_i^{lkg} (\mathbf{B}_i^{lkgT} \mathbf{B}_i^{lkg})^{-1} \mathbf{B}_i^{lkgT} \mathbf{y}. \quad (2.5)$$

The minimum distance between query and feature plane of class i is denoted as

$$r_i(\mathbf{y}) = \min_{l,k,g} \|\mathbf{y} - \mathbf{p}_i^{lkg}\|_2. \quad (2.6)$$

The classification result of NFP follows the rule in (2.2).

For the i th class, the number of constructed feature plane is $K_i = n_i(n_i - 1)(n_i - 2)/6$. The total number of constructed feature planes by all classes is $K = \sum_{i=1}^c K_i$. Compared with NN, although NFP could handle more facial variations, its computational complexity is much higher.

2.1.4 Nearest Feature Subspace

To capture sufficient facial variations without significantly increasing the computational burden, Chien *et al.* [69] generalized the geometrical concept from plane to subspace. As a result, each class only constructs one subspace, which is much less than the number of constructed feature lines and planes. All training samples of a single class span a subspace $\mathbf{S}_i = sp(\mathbf{A}_i)$. The projection of query sample onto subspace $\mathbf{S}_i$ is denoted by $\mathbf{p}_i$, which could be obtained as

$$\mathbf{p}_i = \mathbf{A}_i(\mathbf{A}_i^T \mathbf{A}_i)^{-1} \mathbf{A}_i^T \mathbf{y}. \quad (2.7)$$

Hence, the distance between query image $\mathbf{y}$ and $\mathbf{p}_i$ is calculated as

$$r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{p}_i\|_2. \quad (2.8)$$

The query image is labelled to the class with minimum $r_i(\mathbf{y})$, for $i = 1, \dots, c$.

2.1.5 Linear Regression Classification

It is widely assumed that samples from the same subject with illumination variations lie in a linear subspace [6, 70]. Inspired by this assumption, a linear regression classification (LRC) [71] is proposed for face recognition. Different from other classifiers, which use the extracted features as input, LRC directly employs raw images as

$$\mathbf{y} = \mathbf{A}_i \mathbf{x}_i, \text{ for } i = 1, 2, \dots, c. \quad (2.9)$$

Figure 2.3: Partitioned face image with 4, 6, and 8 modules.

Here, least square is applied to estimate the regression coefficient vector $\mathbf{x}_i$ for each class. The optimum solution is achieved by

$$\mathbf{x}_i = (\mathbf{A}_i^T \mathbf{A}_i)^{-1} \mathbf{A}_i^T \mathbf{y}. \quad (2.10)$$

As a result, the reconstruction image of class i is

$$\mathbf{y}_i = \mathbf{A}_i \mathbf{x}_i = \mathbf{A}_i (\mathbf{A}_i^T \mathbf{A}_i)^{-1} \mathbf{A}_i^T \mathbf{y}. \quad (2.11)$$

The distance between the reconstructed image of class i and the query image is calculated as

$$r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{y}_i\|_2. \quad (2.12)$$

It is not difficult to see that LRC is in fact equivalent to NFS.

To deal with the contiguously occluded query image, Naseem *et al.* [71] employed a modular approach [72] to facilitate LRC. Face images are divided into a number of sub-images. Each training image is partitioned in L modules as shown in Fig. 2.3. Then the l th partitions of training images from the i th class are used to construct a class- and partition-specific subspace $\mathbf{A}_i^l \in \mathbb{R}^{(m/L) \times n_i}$. The query image is also partitioned into L modules, and the l th partition is

denoted as $\mathbf{y}^l \in \mathbb{R}^{m/L}$. Therefore, for the l th modules, $\mathbf{y}^l$ could be regressed by $\mathbf{A}_i^l$ as

$$\mathbf{y}^l = \mathbf{A}_i^l \mathbf{x}_i^l. \quad (2.13)$$

LRC is applied to solve (2.13) and the reconstructed module by $\mathbf{A}_i^l$ is denoted as $\mathbf{y}_i^l$. The distance between the reconstructed module and the query module is calculated as

$$r_i(\mathbf{y}^l) = \|\mathbf{y}^l - \mathbf{y}_i^l\|_2. \quad (2.14)$$

For the l th partition, its decision is denoted as k^l , which is the class with minimum distance:

$$r(\mathbf{y}^l) = \min_i r_i(\mathbf{y}^l). \quad (2.15)$$

As a result, we have L decisions k^l and corresponding distances $r(\mathbf{y}^l)$, for $l = 1, 2, \dots, L$. The query image is classified into the class with minimum module distance, k^{l^*} . Here, l^* is obtained by

$$l^* = \arg \min_l r(\mathbf{y}^l). \quad (2.16)$$

In this thesis, we name this method as modular LRC or LRC with patch (LRC-P).

As the discriminative information of one module is much less than that of the whole image, the results based on a single module is not reliable.

2.1.6 Sparse Representation based Classification

By stacking the training samples of all classes, it generates a dictionary $\mathbf{A}$. The query sample $\mathbf{y}$ can be reconstructed by a linear combination of columns of $\mathbf{A}$ as

$$\mathbf{y} = \mathbf{Ax}. \quad (2.17)$$

Although n_i is usually smaller than m , $n = \sum n_i$ could be greater than m . In this case, (2.17) is an under-determined linear system, which has infinity solutions with error free representation. To achieve an unique, stable, and informative representation coefficient vector $\mathbf{x}$, additional constraint should be applied to (2.17).

As it is widely assumed that face images of an individual class under different illuminations or expression lie in a low-dimensional face subspace [6, 70], the query image can be well reconstructed by the linear combination of sufficient training data of correct subject. Hence, in (2.17), it can expect to only use the training samples from the correct subject to represent $\mathbf{y}$. Consequently, if there are 20 classes, it leads to that 95% elements of $\mathbf{x}$ are zero, which is a sparse representation by all training samples. Therefore, sparse representation based classification (SRC) [67] proposes to constrain the number of nonzero coefficients to be small. To this end, it is rational to employ the ℓ_0 -norm that counts the number of nonzero entries to enforce a sparse vector $\mathbf{x}$ in (2.17). As a result, it derives the following optimization problem

$$\mathbf{x}_0 = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{Ax}. \quad (2.18)$$

However, searching the sparsest solution of Eq. (2.18) is NP-hard. That is, the most efficient method to solve (2.18) is to exhaust all subsets of the coefficients for $\mathbf{x}$. Fortunately, thanks to the studies of compressive sampling [38, 73, 74], the solution of (2.18) can be achieved by

replacing the ℓ_0 -norm with ℓ_1 -norm if $\mathbf{x}_0$ is sparse enough. The ℓ_1 -norm is the nearest convex norm to ℓ_0 -norm. Moreover, ℓ_1 -norm minimization problem can be efficiently solved by the standard linear programming methods.

By substituting the ℓ_0 -norm with ℓ_1 -norm in (2.18), the following ℓ_1 -norm minimization problem is obtained

$$\mathbf{x}_1 = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \quad (2.19)$$

In real images, as the noises cannot be ignored, it is impossible to exactly represent the query image by the training samples of the correct subject. Therefore, the problem (2.19) is converted to a noise tolerated version as [67]

$$\mathbf{x}_1 = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon. \quad (2.20)$$

Many efficient algorithms [75–87] are proposed to solve (2.19) and (2.20). A detailed review and comparison of these algorithms are presented in [88].

Different from the traditional classifiers, the sparse representation scheme lets the training sample compete with the others to win its share in the representation of query image. It is a discriminative progress.

Once achieving the sparse coefficients $\mathbf{x}_1$ via (2.19) or (2.20), the class-specific reconstruction can be generated by using the coefficients associated with a single subject. Let δ_i denote the characteristic function that only retains the coefficients associated with the i th class and sets others to zero. The i th class approximation of $\mathbf{y}$ is defined as

$$\mathbf{y}_i = \mathbf{A}\delta_i(\mathbf{x}_1). \quad (2.21)$$

The query sample $\mathbf{y}$ is assigned to the subject that minimizes the class-specific error between $\mathbf{y}$ and $\mathbf{y}_i$:

$$i^* = \arg \min_i \|\mathbf{y} - \mathbf{y}_i\|_2. \quad (2.22)$$

If an image is contaminated by sparse noise but with arbitrary magnitude, it is impossible to model the errors by ℓ_2 -norm, which is used to capture the small dense Gaussian distributed noises. SRC [67] further considers a corrupted query image as a summation of the original clean image $\mathbf{y}_0$ and sparse error image $\mathbf{e}_0$ as:

$$\mathbf{y} = \mathbf{y}_0 + \mathbf{e}_0 = \mathbf{A}\mathbf{x}_0 + \mathbf{e}_0 = [\mathbf{A}, \mathbf{I}] \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{e}_0 \end{bmatrix} = \mathbf{B}\mathbf{z}, \quad (2.23)$$

where $\mathbf{B} = [\mathbf{A}, \mathbf{I}]$ and $\mathbf{z} = \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{e}_0 \end{bmatrix}$.

The sparsity of both the coefficient vector $\mathbf{x}_0$ and the noise term $\mathbf{e}_0$ results in that $\mathbf{z}$ is also sparse. Therefore, the SRC model can be applied to (2.23) as

$$\mathbf{z}_1 = \arg \min_{\mathbf{z}} \|\mathbf{z}\|_1 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{B}\mathbf{z}\|_2 \leq \epsilon. \quad (2.24)$$

Once the sparse solution $\mathbf{z}_1 = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{e}_1 \end{bmatrix}$ is achieved via ℓ_1 -norm minimization, $\mathbf{y}_r = \mathbf{y} - \mathbf{e}_1$ recovers a clean version of the corrupted query image. To classify the query image, (2.22) is

modified as:

$$i^* = \arg \min_i \|\mathbf{y}_r - \mathbf{A}\delta_i(\mathbf{x}_1)\|_2. \quad (2.25)$$

Fig. 2.4 shows an example, where the occlusion is well approximated by the sparse vector $\mathbf{e}_1$ and the significant nonzero coefficients are associated to the correct class. In this thesis, we name this method as robust SRC (R-SRC) [67].

In some practical scenarios, the occlusion falls on some patches of query image which is prior unknown connected area. For example, if a face is disguised by a scarf, the lower half is almost occluded. In this case, the error term $\mathbf{e}_0$ in (2.23) can not be considered as a sparse vector. Although R-SRC can work well with small ratio of corruption in query image, as the ratio increases, its performance degrades dramatically.

Inspired by [72], Wright *et al.* [67] proposed a modular approach to deal with the large continuous occlusion. Each training image is partitioned into L modules, and it generates a set of modular dictionaries $\mathbf{A}^1, \dots, \mathbf{A}^L \in \mathbb{R}^{(m/L) \times n}$. Similarly, query image $\mathbf{y}$ is partitioned into $\mathbf{y}^1, \dots, \mathbf{y}^L \in \mathbb{R}^{m/L}$. R-SRC is applied to the l th modules of training and query samples, for $l = 1, \dots, L$. Therefore, there are L decisions associated with L modules. The final result is aggregated by voting of all modules. In this thesis, we denote this approach by partitioning as modular SRC or SRC by patch (SRC-P).

2.1.7 Extension of SRC

SRC differentiates itself from the traditional classifiers in three aspects. The first is the utilization of training samples from all classes simultaneously. The second is sparsely representation of query image. The last is the sparsely approximation of the representation errors. Thanks to these merits, SRC achieves some interesting and impressive results in face recognition, which

Figure 2.4: Overview of SRC approach. SRC represents a query image y (a), which is partially occluded, as a sparse linear combination x_1 (b) of all the training sample A (c) plus a sparse error e_1 (d). The red coefficients correspond to training images of the correct individual.

attracts many extension works in recent years.

Many works focus on handling occluded or corrupted query image. In SRC [67], an identity matrix $I \in \mathbb{R}^{m \times m}$ is introduced as the occlusion dictionary appended to training data as (2.23). This scheme significantly increases the computational complexity of SRC. To tackle this problem, a Gabor feature based sparse representation (GSRC) is proposed to reduce the huge computation burden [89]. By using Gabor feature of training data and Gabor occlusion dictionary, GSRC [89] improves the recognition accuracy and significantly reduces the computational cost in dealing with occlusion and corruption. Instead of solving the representation coefficients and sparse errors simultaneously in (2.24), Yang *et al.* optimized x and e separately and iteratively [88]. As a result, the occlusion dictionary is not necessary and a fast implementation is introduced. Some works [90, 91] show doubts about utilization of ℓ_1 -norm to model the noises. In [90], He *et al.* proposed a sparse representation algorithm that measures the contamination based on the maximum correntropy criterion [92], which is much more insensitive to outliers. On the other hand, by analyzing the Laplace distribution (i.e., ℓ_1 -norm) and Gaussian distribution (i.e., ℓ_2 -norm), Yang *et al.* [91] argued that both of them are not optimum for face with occlusion or disguise. In [91], a regularized robust coding (RRC) approach

is proposed to model the errors based on the maximum a posterior principle. There are some recently proposed methods [93–95] attempting to cooperate the sparse representation with the prior knowledge of corruption. In the reality, the error pixels often form a continuous region. To identify the occluded pixels with this prior knowledge, Zhou *et al.* [93] proposed to integrate a Markov Random Field into the sparse coding. The occlusion part of query is detected and removed from query and training samples, and then the rest parts are used for sparse representation. In [94, 95], Deng *et al.* showed that the variation (e.g., illumination, expression, and disguise) of a single subject can be shared across all subjects. As a result, an intra-class variation dictionary is generated by subtracting a natural or prototype or centroid image of each class from training images of the same subject [94, 95]. The disguise of the query image can be represented as a sparse combination of the intra-class variations [94, 95].

As SRC aims at finding the most compact representation via minimizing ℓ_1 -norm of coefficients, it may not be the best selection for classification. To achieve a more discriminative representation or coefficients of the query image, many algorithms [96–100] are introduced. In SRC, the label information of training data is ignored, which is crucial for recognition task. To boost the performances of face recognition, Majumdar *et al.* [96] proposed to seek a sparse representation in terms of the class number instead of the sample number [101]. In addition, local structure is useful in real applications [102]. In [103], as locality must lead to sparsity but not necessarily vice versa, Yu *et al.* claimed that locality is more essential than sparsity. Therefore, Wang *et al.* [97] replaced the sparse constraint with the locality restriction, which still leads to sparse coefficients. By utilizing both the group sparsity and locality constraint, Chao *et al.* [99] proposed a locality and group sensitive sparse representation (LGSR) to achieve better performances. It is well known that nonnegative matrix factorization [104, 105] also leads to a sparse representation. Moreover, the nonnegativity representation makes it easy to interpret thanks to purely additive combinations of basis vectors. Therefore, in addition to the sparse constraint, Liu *et al.* [98] enforced nonnegative coefficients. Although many works give the credit of SRC's

success to the sparse representation coefficients, Zhang *et al.* [100] argued that the collaborative representation is more important than the sparsity constraint. Zhang *et al.* [100] replaced the ℓ_1 -norm over coefficients with ℓ_2 -norm, which achieves competitive results with SRC but requires less computational time.

In [67], SRC directly uses the training data (i.e., $\mathbf{A}$) as the dictionary, which is the simplest but not the optimum for sparse representation and classification. To tackle this problem, many dictionary learning algorithms are proposed. Aharon *et al.* [106] introduced an algorithm named K-SVD to learn a dictionary such that it can well represent each training sample under strict sparse constraint. Although K-SVD shows impressive results on image reconstruction, it may be not capable for classification. By taking both representation and discrimination power into consideration, a discriminative K-SVD is proposed in [107]. Label Consistent-KSVD (LC-KSVD) [108, 109] further proposes a label consistent constraint, which is appended to the reconstruction and classification error. As the above dictionary learning methods attempt to find an optimum dictionary for all subjects, they cannot use the class-specific reconstruction error for classification as SRC (2.22). Some recently works [110–113] focus on learning class-specific dictionary. Mairal *et al.* [110] proposed to learn the different dictionaries for different subjects. That is, there is one dictionary for each subject. In [110], it introduces a discriminative term based on the idea that the dictionary should be good at representing the samples of the same class and bad for the others. To differentiate the dictionaries between classes, Ramirez *et al.* [111] minimized the coherence between dictionaries. Yang *et al.* [112, 113] proposed a Fisher discriminative dictionary learning (FDDL). In addition to utilization of the representation error, FDDL constrains the representation coefficients of the same class to be similar and those of different classes to be dissimilar. One common weakness of all previously discussed approaches is that they do not work well on the corrupted training images. To alleviate this problem, some works [114–116] based on low rank decomposition or robust principle component analysis (RPCA) [117] are proposed. Chen *et al.* [114] learned one low-rank dictionary for

each class and minimized the coherence between dictionaries for increasing their discriminative power. Ma *et al.* [115] proposed an objective function combining sparse coefficients, class discrimination and rank minimization for dictionary learning. Following the idea of subspace clustering by low-rank representation [118], Zhang *et al.* [116] proposed a low-rank representation for image classification. To improve the performances, [116] enforces the training samples of same class having the same representation. Although the low-rank procedure can alleviate the problems of corrupted data, it also decreases the representation power of each class.

Recently, SRC is applied to feature extraction or dimension reduction. In [119], Qiao *et al.* proposed a sparsity preserving projections (SPP) to learn the sparse structure of training data and preserve the learnt structure in the low dimensional subspace. As SPP is an unsupervised approach, it may not be the best selection for extract the most discriminative feature. In [120], Gui *et al.* captured the underlying sparsity structure within the correct class. However, the sparse constraint within class is not necessary. Several works [121, 122] propose to simultaneously learn a discriminative feature and dictionary for sparse representation within one framework.

2.2 Linear Feature Extraction in Face Recognition

Feature extraction is another important problem of face recognition. The facial images are usually with high dimensionality (e.g., 480000 for a 600×800 image). If we directly process the original image data, an extremely high computational burden is required. Further more, the data with such high dimensionality is potential to suffer the so-called "curse of dimension" problem [123]. As the training data does not contain all kinds but only a particular type of image (i.e., human faces), it should be resided within a much lower dimensional subspace [6, 70]. To explore the underlying low dimensional structure of the original data, numerous feature

extraction methods [1, 6, 119, 124–133] are proposed. Some related approaches are reviewed in the following subsections.

2.2.1 Principle Component Analysis

In [124, 125], principle component analysis (PCA) is applied to face recognition for maximizing the variance of data in the subspace. Towards this end, the following optimization problem is proposed

$$\mathbf{W}^* = \arg \max_{\mathbf{W}} \sum_{k=1}^n \|\mathbf{W}^T(\mathbf{a}_k - \mathbf{m})\|_2^2, \quad (2.26)$$

where $\mathbf{a}_k$ is the k th training sample, and $\mathbf{m}$ is the mean face of all training samples. The solution of (2.26) can be obtained as $\mathbf{W}^* = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d]$, which consists of d eigenvectors corresponding to the d largest eigenvalues of the total covariance matrix $\mathbf{S}_t$. Here,

$$\mathbf{S}_t = \frac{1}{n} \sum_{k=1}^n (\mathbf{a}_k - \mathbf{m})^T (\mathbf{a}_k - \mathbf{m}). \quad (2.27)$$

Once the optimized projection matrix $\mathbf{W}^*$ is attained, the feature of each samples is generated as $\tilde{\mathbf{a}}_k = \mathbf{W}^{*T} \mathbf{a}_k$. PCA, which targets for the optimum of reconstruction, is not the best choice for classification.

2.2.2 Linear Discriminative Analysis

In PCA, it treats all training samples equally by ignoring their identities, which is crucial for classification task. To overcome this problem, linear discriminative analysis (LDA) [6, 127] is introduced to minimize within class distances and maximize between classes distances in the learnt low dimensional subspaces. It defines a within class covariance matrix $\mathbf{S}_w$ and between

class covariance matrix $\mathbf{S}_b$ as

$$\mathbf{S}_w = \sum_{i=1}^c \sum_{j=1}^{n_i} \frac{1}{n_i} (\mathbf{a}_{i,j} - \mathbf{m}_i)^T (\mathbf{a}_{i,j} - \mathbf{m}_i), \quad (2.28)$$

$$\mathbf{S}_b = \sum_{i=1}^c \frac{1}{n_i} (\mathbf{m}_i - \mathbf{m})^T (\mathbf{m}_i - \mathbf{m}), \quad (2.29)$$

here, $\mathbf{m}_i$ is the centroid of the i th class. $\mathbf{S}_w$ and $\mathbf{S}_b$ capture within class distances and between class distances, respectively. LDA proposes to minimize $\mathbf{S}_w$ and maximize $\mathbf{S}_b$ simultaneously as

$$\mathbf{W}^* = \arg \max_{\mathbf{W}} \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}}. \quad (2.30)$$

The optimum $\mathbf{W}^*$ can be obtained via solving the generalized eigenvalue decomposition problem

$$\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}. \quad (2.31)$$

$\mathbf{W}^* = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d]$ are the d eigenvectors corresponding to the d largest eigenvalues. The same as PCA, LDA feature of $\tilde{\mathbf{a}}_k$ is generated as $\mathbf{W}^{*T} \mathbf{a}_k$.

As the rank of $\mathbf{S}_b$ is upper bounded by $c - 1$, d is less than or equal to $c - 1$.

2.2.3 Neighborhood Preserving Embedding

Different from PCA and LDA, neighborhood preserving embedding (NPE) [130] proposes to preserve the local structure in the subspace. For each training sample $\mathbf{a}_k$, its K nearest neighborhood are denoted as $\pi_K(\mathbf{a}_k)$. The local structure is captured by an affinity matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$

whose k th column is obtained by

$$\begin{aligned} & \min \|\mathbf{a}_k - \sum_{j \in \pi_K(\mathbf{a}_k)} X_{k,j} \mathbf{a}_j\|_2^2 \\ \text{s.t. } & \sum_{j \in \pi_K(\mathbf{a}_k)} X_{k,j} = 1, \quad X_{k,j} = 0 \text{ if } j \notin \pi_K(\mathbf{a}_k). \end{aligned} \quad (2.32)$$

To preserve the neighborhood structure in the subspace, a reasonable criterion of choosing an optimal projection matrix $\mathbf{W}$ is to optimize the cost function as following:

$$\mathbf{W}^* = \arg \min_{\mathbf{W}} \sum_{k=1}^n \|\mathbf{W}^T(\mathbf{a}_k - \sum_{j \in \pi_K(\mathbf{a}_k)} X_{k,j} \mathbf{a}_j)\|_2^2 \quad \text{s.t.} \quad \mathbf{W}^T \mathbf{A} \mathbf{A}^T \mathbf{W} = \mathbf{I}, \quad (2.33)$$

here, $\mathbf{W}^T \mathbf{A} \mathbf{A}^T \mathbf{W} = \mathbf{I}$ is imposed to remove an arbitrary scaling factor. The above optimization problem can be solved via generalized eigenvalue decomposition as

$$\mathbf{A} \mathbf{M}_N \mathbf{A}^T \mathbf{w} = \lambda \mathbf{A} \mathbf{A}^T \mathbf{w}, \quad (2.34)$$

where $\mathbf{M}_N = (\mathbf{I} - \mathbf{X})^T (\mathbf{I} - \mathbf{X})$. The solution of Eq. (2.33) are the d eigenvectors corresponding the smallest d eigenvalues of Eq. (2.34).

2.2.4 Sparsity Preserving Projections

As sparse representation shows its discriminative power for classification, Qiao *et al.* [119] proposed a sparsity preserving projections (SPP) to maintain these sparsity structure of training data in the subspace. The affine weight matrix $\mathbf{X}$ of SPP is learnt by the sparse representation of training data. For the k th column of $\mathbf{X}$, i.e. $\mathbf{x}_k$, is computed as following optimization problem

$$\mathbf{x}_k = \arg \min_{\mathbf{x}} \|\mathbf{a}_k - \mathbf{A} \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \sum_{j=1}^n x_j = 1 \text{ and } x_k = 0. \quad (2.35)$$

where $\mathbf{x} = [x_1, \dots, x_j, \dots, x_n]$.

To preserve this sparsity structure in the subspace, the optimal projection matrix $\mathbf{W}^*$ is obtained by optimizing the following cost function:

$$\mathbf{W}^* = \arg \min_{\mathbf{W}} \sum_{i=k}^n \|\mathbf{W}^T(\mathbf{a}_k - \mathbf{A}\mathbf{x}_k)\|_2^2 \quad \text{s.t.} \quad \mathbf{W}^T \mathbf{A} \mathbf{A}^T \mathbf{W} = \mathbf{I}. \quad (2.36)$$

The above optimization problem can be solved via generalized eigenvalue decomposition as

$$\mathbf{A} \mathbf{M}_S \mathbf{A}^T \mathbf{w} = \lambda \mathbf{A} \mathbf{A}^T \mathbf{w}, \quad (2.37)$$

where $\mathbf{M}_S = (\mathbf{I} - \mathbf{X})^T (\mathbf{I} - \mathbf{X})$. Similar to NPE, the solution of Eq.(2.36) are the d eigenvectors corresponding the smallest d eigenvalues of Eq.(2.37).

2.3 Summary

Although the existing sparse representation based approaches show their superiority over the traditional face recognition ones in some challenging scenarios, there are still several problems needed to be addressed:

- In [119], Qiao *et al.* proposed a sparsity preserving feature extraction method for face recognition. As sparse representation is an unsupervised procedure, it may not capture sufficient discrimination. It is well known that label information of training data is critical for classification. To tackle this issue, we present an algorithm of supervised sparse representation and apply it to dimensional reduction in Chapter 3.
- One disadvantage of modular SRC is that it treats all modules equally and does not regard whether it is clean or occluded. Moreover, it searches for multiple local optimums, each

of which is based on a single module. However, a single module has less discriminative information for classification than the whole image, which will lead to an unreliable local optimum. The fusion on these unreliable results may cause a misclassification. To alleviate these problems of modular SRC, we proposed a modular weighted global sparse representation and a collaborative patch representation in Chapter 4 and Chapter 5, respectively.

- If the data originating from distinct classes are independent, SRC can achieve a desired representation with the significant coefficients only associated to the correct subject. However, the face subspaces spanned by different subjects are correlated because of consisting of the same facial components (e.g., eyes, nose, and mouth). Therefore, SRC may represent a query face by training samples of many different subjects. Moreover, to achieve a sparse representation, few samples will be randomly selected from the highly correlated training samples, which may lead to a wrong classification. We proposed a class-wise sparse representation to overcome these limitations of SRC in Chapter 5.
- With sufficient representative and carefully controlled training data, SRC achieves some impressive results. Face identification is an example where we have a large number of identities but sufficient representative and uncorrupted training images cannot be guaranteed for every subject. A violation of these two conditions leads to a poor performance of SRC. To address this critic issue, we propose a sparse- and dense-hybrid representation framework facilitated by the proposed supervised low-rank dictionary decomposition in Chapter 6.

Chapter 3

Discriminative Sparsity Preserving Embedding for Face Recognition

The coefficients achieved by sparse representation show some discriminative information. However, as an unsupervised approach, sparse representation may not be the optimum for learning the structure of labelled training data. To address this problem, we propose an algorithm, namely supervised sparse representation (SSR), to cooperate with the class label information of training data. Furthermore, we also propose a feature extraction or dimension reduction method, named discriminative sparsity preserving embedding (DSPE). The proposed DSPE captures the discriminative sparse structure of training data by employing SSR and finds the low dimensional subspace that reduces the within class distances meanwhile keeping the between class distances.

3.1 Introduction

The facial images are usually with high dimensionality (e.g., 480000 for a 600×800 image). To directly process the image data, it requires very high computational complexity. Moreover, it is potential to suffer the so-called "curse of dimension" problem [123]. Fortunately, face

images are proved to reside within a low dimensional subspace [6, 70]. In order to find the underlying structure of the original data, numerous dimensionality reduction (DR) methods [125, 127, 130, 132, 134] are proposed.

As discussed in Chapter 2, principle component analysis (PCA), linear discriminant analysis (LDA) and locality preserving projections (LPP) are proposed to optimize different criteria. These methods have yield reasonable good performance on some benchmark face databases. Because these approaches utilize the original training data as isolated sample points, they are not able to learn the geometrical structure with limited number of training samples. To tackle this problem, neighborhood preserving embedding (NPE) [130] was proposed to extend the limited sample points to manifolds. It aims at preserving the local manifold structure constructed by the k nearest neighbors. Inspired by sparse representation (SR) [67, 135, 136] and NPE, sparsity preserving projections (SPP) [119] constructs a sparse weight matrix which stores the sparse structure and finds the low dimensional embedding to best preserve such structure. Due to the natural discriminating power of sparsest representation, SPP is able to find a discriminative mapping.

Although NPE and SPP can capture some discriminating power, as unsupervised methods, they are not optimal for classification. In this chapter, we propose an algorithm, namely supervised sparse representation (SSR), to explore the sparse structure cooperating with label information. Furthermore, we propose a DR method, discriminative sparsity preserving embedding (DSPE). DSPE learns the discriminative structure with SSR, and finds the low dimensional subspace that reduces the within class distances and keeps the between class distances.

3.2 Proposed Framework for Feature Extraction

Let $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_i, \dots, \mathbf{a}_n]$, where $\mathbf{a}_i \in \mathbb{R}^m$ is the feature vector representing the i th training image and n is the total number of training images. The class or subject label of $\mathbf{a}_i$ is denoted as $l(i) \in \{1, 2, \dots, c\}$, here c is the number of subjects.

We consider the i th training sample being represented as a sparse linear combination of the rest training samples [67] as

$$\mathbf{a}_i = \mathbf{A}\mathbf{x}_i + \mathbf{e}, \quad (3.1)$$

where $\mathbf{x}_i = [x_{i,1}, \dots, x_{i,i-1}, 0, x_{i,i+1}, \dots, x_{i,n}]^T$ and $\mathbf{a}_i$ is assumed to be coupled with additive independent and identically distributed (i.i.d.) Gaussian noise $\mathbf{e}$ with zero-mean and σ standard variation.

To obtain a sparse representation, element of coefficient $\mathbf{x}_i$ follows the i.i.d. Laplace distribution [137], that is

$$p(x_{i,j}) = \frac{\lambda}{2} e^{(-\lambda|x_{i,j}|)} \quad \text{for } i \neq j, \quad (3.2)$$

where λ is the scale parameter.

The argument achieving the maximum a posteriori estimation of $p(\mathbf{x}_i|\mathbf{a}_i, \mathbf{A})$ is derived as

$$\begin{aligned} & \underset{\mathbf{x}_i}{\operatorname{argmax}} \quad p(\mathbf{a}_i|\mathbf{A}, \mathbf{x}_i)p(\mathbf{x}_i) \\ &= \underset{\mathbf{x}_i}{\operatorname{argmax}} \quad \frac{1}{(2\pi\sigma^2)^{n/2}} e^{(\frac{\|\mathbf{a}_i - \mathbf{A}\mathbf{x}_i\|_2^2}{-2\sigma^2})} \prod_{\substack{j=1 \\ j \neq i}}^n \frac{\lambda}{2} e^{(-\lambda|x_{i,j}|)} \\ & \quad \text{s.t. } x_{i,i} = 0. \end{aligned} \quad (3.3)$$

The penalized log likelihood of Eq. (3.3) has the form

$$\begin{aligned}
& \underset{\mathbf{x}_i}{\operatorname{argmax}} \log p(\mathbf{a}_i | \mathbf{A}, \mathbf{x}_i) p(\mathbf{x}_i) \\
&= \underset{\mathbf{x}_i}{\operatorname{argmax}} -\frac{1}{2\sigma^2} \|\mathbf{a}_i - \mathbf{A}\mathbf{x}_i\|_2^2 - \sum_{j=1}^n \lambda |x_{i,j}| \\
&= \underset{\mathbf{x}_i}{\operatorname{argmin}} \|\mathbf{a}_i - \mathbf{A}\mathbf{x}_i\|_2^2 + \sum_{j=1}^n 2\sigma^2 \lambda |x_{i,j}| \\
&= \underset{\mathbf{x}_i}{\operatorname{argmin}} \|\mathbf{a}_i - \mathbf{A}\mathbf{x}_i\|_2^2 + \sum_{j=1}^n \gamma |x_{i,j}| \\
&= \underset{\mathbf{x}_i}{\operatorname{argmin}} \|\mathbf{a}_i - \mathbf{A}\mathbf{x}_i\|_2^2 + \gamma \|\mathbf{x}_i\|_1 \\
&\quad \text{s.t. } x_{i,i} = 0,
\end{aligned} \tag{3.4}$$

where $\gamma = 2\sigma^2\lambda$. Eq. (3.4) is the original SR and employed in SPP to learn the sparse structure.

For the classification task, the label information of training sample is very crucial. However, in the original SR, all coefficients are following the same distribution as (3.2). Therefore, the label information does not cooperates with SR.

3.2.1 Supervised Sparse Representation

In many practical applications, the label of training samples L can be attained. We consider the prior distribution of element of coefficient $\mathbf{x}_i$ being conditional with label L . For the coefficients associating with training samples from the same subject, it is assumed that they follow the same distribution. Therefore, the probability density function (pdf) $p(x_{i,j})$ in (3.2) should be converted to a class conditional version $p(x_{i,j} | L)$ as

$$p(x_{i,j} | L) = \frac{\lambda_{l(j)}^i}{2} e^{(-\lambda_{l(j)}^i |x_{i,j}|)} \quad \text{for } i \neq j. \tag{3.5}$$

Similar to Eq. (3.3) and (3.4), to calculate the optimal argument $\mathbf{x}_i$ achieving the maximum a posteriori estimation of $p(\mathbf{x}_i|\mathbf{a}_i, \mathbf{A}, L)$, the proposed supervised sparse representation (SSR) is formulated as

$$\underset{\mathbf{x}_i}{\operatorname{argmin}} \|\mathbf{a}_i - \mathbf{A}\mathbf{x}_i\|_2^2 + \beta \|\Lambda^i \mathbf{x}_i\|_1 \quad \text{s.t. } x_{i,i} = 0, \quad (3.6)$$

where $\beta = 2\sigma^2$, Λ^i is a diagonal matrix and $\Lambda_{jj}^i = \lambda_{l(j)}^i$.

As label information is ignored, SPP [119] cannot guarantee to well capture the discriminative information. That is, there are few or no nonzero coefficients associating with the correct subject. In this case, the sparse structure of SPP is useless, or even harmful for classification. Therefore, to build a more efficient method, more nonzero coefficients should be associating with the correct subject.

In Fig. 3.1, the pdf of Laplace distribution with various parameters λ are plot. We can see that, as the λ decreases, the probability of $x_{i,j}$ falling in the region around zero (the area between the pdf and the dash lines) also decreases. Hence, the probability of $x_{i,j}$ being nonzero increases as λ decreases. To encourage more nonzero coefficients from the correct subject, obviously, the coefficients associating with the correct subject should be assigned with small λ . To suppress other samples, their coefficients should be assigned with large λ . As a result, if the j th training sample from the same subject as the i th sample, that is $l(i) = l(j)$, Λ_{jj}^i is set to a positive value $\theta < 1$. Otherwise, Λ_{jj}^i is fixed as one. Thus, the diagonal matrix Λ^i is given by

$$\Lambda_{jj}^i = \begin{cases} \theta, & l(i) = l(j) \\ 1, & \text{otherwise.} \end{cases} \quad (3.7)$$

Denote $\mathbf{z}_i = \Lambda^i \mathbf{x}_i$. As Λ^i is with all positive diagonal elements, Λ^i is invertible. Substituting

Figure 3.1: pdf of Laplace distribution with different selections of λ .

$\mathbf{x}_i = \mathbf{\Lambda}^{i-1} \mathbf{z}_i$ into Eq. (3.6) yields

$$\begin{aligned} \mathbf{z}_i &= \arg \min_{\mathbf{z}} \|\mathbf{a}_i - \mathbf{A} \mathbf{\Lambda}^{i-1} \mathbf{z}\|_2^2 + \beta \|\mathbf{z}\|_1 \\ &= \arg \min_{\mathbf{z}} \|\mathbf{a}_i - \mathbf{B} \mathbf{z}\|_2^2 + \beta \|\mathbf{z}\|_1 \\ \text{s.t. } z_i &= 0, \end{aligned} \quad (3.8)$$

where $\mathbf{z} = [z_1, \dots, z_i, \dots, z_n]$ and $\mathbf{B} = \mathbf{A} \mathbf{\Lambda}^{i-1}$. Eq. (3.8) is a ℓ_1 minimization problem. Once $\mathbf{z}_i$ is generated, the optimum solution of Eq. (3.6) can be attained by $\mathbf{x}_i = \mathbf{\Lambda}^{i-1} \mathbf{z}_i$.

After generating $\mathbf{x}_i$ for each training sample, for $i = 1, \dots, n$, the discriminative sparse reconstructive weight matrix $\mathbf{X}$ of the proposed discriminative sparsity preserving embedding (DSPE) method is obtained as follows:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T. \quad (3.9)$$

To evaluate the discriminative power of different methods, the Yale face database [6] is used. Six images per subject are randomly chosen to learn the structure matrices. Fig. 3.2 shows the sub-block of distribution of nonzero coefficients lying in matrices captured by NPE, SPP and DSPE. For NPE, the neighbor parameter k is set to five. The nonzero coefficients in the matrices are mark as plus signs. Let i denote the number on the vertical axis and j denote the number on the horizontal axis. The plus sign located at (i, j) means that the j th sample $\mathbf{a}_j$ is used to represent the i th sample $\mathbf{a}_i$. Element (i, j) residing in the red square means that $\mathbf{a}_i$ and $\mathbf{a}_j$ share the same class label. In Fig. 3.2a, only the k nearest neighbors are selected to reconstruct the target sample. As the within-class variation can be larger than the between class variation due to illumination, pose, and expression, many selected samples are not from the correct subject. Because of the natural discriminating power of SR, SPP tends to learn the discriminative structure. Therefore, Fig. 3.2b has more plus signs residing in the red squares

than that of Fig. 3.2a. It should be noted that, in some rows of Fig. 3.2a and b, no samples from the correct subjects are found (no plus sign residing within the red square). This structure is harmful for classification. It is clear that, for all the rows of Fig. 3.2c, there are more nonzero coefficients from the correct subject than that of Fig. 3.2a and b. In addition, in each row, more than one samples from the correct subject are presented to represent the target samples. As a result, the discriminating power explored by DSPE is superior to those of NPE and SPP.

3.2.2 Preserving the Supervised Sparse Reconstructive Coefficients

After obtaining the discriminative sparse reconstructive weight matrix $\mathbf{X}$, we partition it into two matrices. One is the correct-class or within-class matrix $\mathbf{X}^w$, which stores the relation of training samples from the correct subjects. The other one is the wrong-class or between-class matrix $\mathbf{X}^b$, which stores the relation of training samples from other subjects. These two matrices are generated by

$$X_{ij}^w = \begin{cases} X_{ij}, & \text{if } l(i) = l(j) \\ 0, & \text{otherwise,} \end{cases} \quad (3.10)$$

$$X_{ij}^b = \begin{cases} X_{ij}, & \text{if } l(i) \neq l(j) \\ 0, & \text{otherwise.} \end{cases} \quad (3.11)$$

The correct-class matrix $\mathbf{X}^w$ reflects the intra-class relationship. We expect to preserve the

Figure 3.2: Distribution of nonzero coefficients in weighting matrices: (a) matrix of NPE, (b) matrix of SPP, (c) matrix of DSPE.

structure in the low dimensional subspace by minimizing the following distance

$$\begin{aligned}
& \min_{\mathbf{w}} \sum_{i=1}^n \|\mathbf{w}^T \mathbf{a}_i - \mathbf{w}^T \mathbf{A} \mathbf{x}_i^w\|_2^2 \\
&= \min_{\mathbf{w}} \mathbf{w}^T \mathbf{A} (\mathbf{I} - \mathbf{X}^w)^T (\mathbf{I} - \mathbf{X}^w) \mathbf{A}^T \mathbf{w} \\
&= \min_{\mathbf{w}} \mathbf{w}^T \mathbf{A} \mathbf{M}^w \mathbf{A}^T \mathbf{w},
\end{aligned} \tag{3.12}$$

where $\mathbf{x}_i^w$ is the i th row of $\mathbf{X}^w$, and $\mathbf{M}^w = (\mathbf{I} - \mathbf{X}^w)^T (\mathbf{I} - \mathbf{X}^w)$.

The wrong-class matrix $\mathbf{X}^b$ characterizes the interclass relation. Different from the correct class, whose reconstructed error should be minimized, the residual of the wrong classes should be large in the learnt subspace. As a result, we maximize its projection in the subspaces, which leads to samples from different classes away. To this end, the following distance is maximized:

$$\begin{aligned}
& \max_{\mathbf{w}} \sum_{i=1}^n \|\mathbf{w}^T \mathbf{a}_i - \mathbf{w}^T \mathbf{A} \mathbf{x}_i^b\|_2^2 \\
&= \max_{\mathbf{w}} \mathbf{w}^T \mathbf{A} (\mathbf{I} - \mathbf{X}^b)^T (\mathbf{I} - \mathbf{X}^b) \mathbf{A}^T \mathbf{w} \\
&= \max_{\mathbf{w}} \mathbf{w}^T \mathbf{A} \mathbf{M}^b \mathbf{A}^T \mathbf{w},
\end{aligned} \tag{3.13}$$

where $\mathbf{x}_i^b$ is the i th row of $\mathbf{X}^b$, and $\mathbf{M}^b = (\mathbf{I} - \mathbf{X}^b)^T (\mathbf{I} - \mathbf{X}^b)$.

To minimize the distances between the correct class samples while maximize the distances between the wrong class samples, Eq. (3.12) and Eq. (3.13) are combined into one optimization problem as

$$\begin{aligned}
& \min_{\mathbf{w}} \mathbf{w}^T \mathbf{A} \mathbf{M}^w \mathbf{A}^T \mathbf{w} - \alpha \mathbf{w}^T \mathbf{A} \mathbf{M}^b \mathbf{A}^T \mathbf{w} \\
&= \min_{\mathbf{w}} \mathbf{w}^T \mathbf{A} (\mathbf{M}^w - \alpha \mathbf{M}^b) \mathbf{A}^T \mathbf{w},
\end{aligned} \tag{3.14}$$

where $\alpha \geq 0$ is a parameter balancing two terms in the objective function.

To avoid an arbitrary scaling factor in the projection, we impose a constraint as follows:

$$\mathbf{w}^T \mathbf{A} \mathbf{A}^T \mathbf{w} = 1. \quad (3.15)$$

Thus, the objective function can be recast as

$$\begin{aligned} & \min_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{A} (\mathbf{M}^w - \alpha \mathbf{M}^b) \mathbf{A}^T \mathbf{w}}{\mathbf{w}^T \mathbf{A} \mathbf{A}^T \mathbf{w}} \\ &= \min_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{A} ((1 - \alpha) \mathbf{I} - \mathbf{M}) \mathbf{A}^T \mathbf{w}}{\mathbf{w}^T \mathbf{A} \mathbf{A}^T \mathbf{w}} \\ &= \max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{A} \mathbf{M} \mathbf{A}^T \mathbf{w}}{\mathbf{w}^T \mathbf{A} \mathbf{A}^T \mathbf{w}}, \end{aligned} \quad (3.16)$$

where $\mathbf{M} = \mathbf{M}^w + \mathbf{M}^{wT} - \mathbf{M}^{wT} \mathbf{M}^w - \alpha (\mathbf{M}^b + \mathbf{M}^{bT} - \mathbf{M}^{bT} \mathbf{M}^b)$.

The optimal $\mathbf{w}$'s are the eigenvectors corresponding to the largest d eigenvalues of the following generalized eigenvalue problem:

$$\mathbf{A} \mathbf{M} \mathbf{A}^T \mathbf{w} = \tau \mathbf{A} \mathbf{A}^T \mathbf{w}. \quad (3.17)$$

The following Algorithm 1 summarizes the proposed DSPE method.

Algorithm 1: Discriminative Sparsity Preserving Embedding (DSPE)

Step 1: Construct the discriminative sparse weight matrix $\mathbf{X}$ using (3.6) and (3.9).

Step 2: Generate the d projection vectors using (3.17).

3.3 Experiments

We evaluate the performance of the proposed DSPE on ORL [138], Yale [6], AR [139], CMU PIE [140]. These face databases focus on frontal faces with illumination and pose variations, disguises and expression changes. In the experiments, we compare the proposed method with

the related methods: LDA with regularization [134], supervised LPP (S-LPP) [132], supervised NPE(S-NPE) [130], SPP [119]. A nearest neighbor classifier with cosine distance is employed in the projected feature space. For all evaluated approaches, the PCA of training data retaining 99% energy is used as input. The parameter α in (3.14) is fixed to 0.05. For the parameter θ in (3.7), we set it to 0.1 based on the experimental evaluation in Section 3.3.5.

3.3.1 Yale Face Database

The Yale database consists of 165 frontal-face images of 15 subjects, captured under various facial expressions, lighting condition and facial detail (glasses or no glasses). For computational effectiveness, all images are resized to 32×32 . To test the effect of the number of image for each individual on the recognition accuracy, we randomly select $t(= 3, 4, 5, 6)$ images of each subject for training and the rest are used for testing. For a given t , the experiments were repeated 20 times. Table 3.1 shows the maximal average face recognition rates for different t of each method.

As it can be seen, the supervised methods (LDA, S-LPP, S-NPE, DSPE) greatly outperform the SPP. For this dataset, SPP cannot capture discriminating power well. Therefore, compared with supervised methods, the recognition rate of SPP is much lower. The proposed DSPE consistently achieves the best results among all methods.

Table 3.1: Recognition Rate on the Yale Face Database.

Methods	LDA	S-LPP	S-NPE	SPP	DSPE
3 train	60.7%	65.2%	65.8%	50.2%	66.1%
4 train	68.6%	73%	73%	57.6%	74.6%
5 train	71.4%	74.4%	74%	58.8%	75.9%
6 train	79.9%	80.7%	80.5%	63%	83.2%

3.3.2 ORL Face Database

The ORL face database contains a total of 400 faces images from 40 subjects (10 samples per person). The images were taken with variation including expression, facial detail, and pose. All images are reduced into the size of 32×32 . For each subject, $t(= 3, 4, 5, 6)$ images are randomly chosen as training set and the rest images are used as testing set. For each t , we repeated the experiment 20 times. The maximal average recognition rates on the testing set against the number of training sample are given in Table 3.2.

The same as the previous subsection, SPP can not compete with the supervised methods. However, for this dataset, SPP learns the discriminative sparse structure much better. As a result, LDA, S-LPP, S-NPE are only slightly superior to SPP. The recognition rate of DSPE are higher than those of the other methods in this experiment.

Table 3.2: Recognition Rate on the ORL Face Database.

Methods	LDA	S-LPP	S-NPE	SPP	DSPE
3 train	84.6%	84.3%	84.7%	83.1%	86%
4 train	89.9%	89.2%	89.4%	88.1%	90.5%
5 train	92.7%	92%	92.7%	90.5%	93.1%
6 train	94.5%	93.6%	94%	92.9%	95.4%

3.3.3 AR Database

We use a subset including 50 male subjects and 50 female subjects of AR database. For each subject, 26 face images are taken in two separate sessions, that is, each session has 13 samples. These two sessions are with the same expression, illumination and disguise variation. However, Session 2 is captured one week later than Session 1. All images are resized from 165×120 to 55×40 . We use 7 undisguised images from Session 1 for training, and 7 undisguised ones from

Figure 3.3: Recognition rates on AR database versus number of dimensionality.

Session 2 for testing.

Fig. 3.3 shows the performances of various methods with different dimensionality. It is not surprise that, as the only unsupervised approach, SPP underperforms the others. The accuracies of LDA, S-LPP and S-NPE are similar. The proposed DSPE consistently performs as the best when the dimension of subspace exceeds 15. The highest recognition rates of DSPE is 92.9%.

3.3.4 CMU PIE Database

In CMU PIE database, there are more than 40,000 facial images captured from 68 subjects. The images were acquired by 13 synchronized cameras and 21 flashes, with varying pose, illumination and expression. In the experiment, we choose the images captured from the frontal poses (C27) and use all the images under different illuminations and expressions. As a result, for each

of 68 subjects, there are 49 facial images. All images are with the size of 45×45 . We randomly select t (10, 20, 30) out of 49 samples per subject for training and the rest are used for testing.

The averaged results over 10 runs are plot in Fig. 3.4. It shows that SPP again achieves the worst result, though the gap between it and the others decreases as the number of training samples increases. When $t = 20$ or 30, the performances of S-LPP and LDA are so close that we cannot distinguish them in the Fig. 3.4b and c. The proposed DSPE always achieves the best recognition rates with different selections of t . The performance gain increases as t decreases.

3.3.5 Effect of θ

The parameter θ in (3.7) controls the weight of the correct subject. If $\theta = 1$, the proposed supervised sparse representation is identical to the original sparse representation. If θ is set to a value less than 1, the correct subject tends to win more share in the representation. Therefore, the value of θ is crucial. In this experiment, we investigate the effect of θ by varying its value from 1 to 0.001 and recording corresponding recognition rates.

Fig. 3.5 plots the recognition rates with different selections of θ on CMU PIE database with 10 training samples per subject. It can be seen that the optimum results are stably achieved by the θ ranging from 0.1 to 0.01. Although $\theta = 1$ (i.e., sparse representation) can learn some discriminative information, it still underperforms the optimum θ with a gap about 2%. On the other hand, when $\alpha = 0.001$, the result is not the optimum either. This may be because, with a very small θ , Eq. (3.7) only selects the coefficients of the correct subject and cannot capture the coefficients associated with the other classes, which results in that the proposed method cannot keep the distances of interclass.

Figure 3.4: Recognition rates on CMU PIE database. The results of 10, 20 and 30 training samples per subject are shown in (a), (b) and (c), respectively.

Figure 3.5: Recognition rates versus θ on CMU PIE database.

3.4 Summary

In this chapter, a supervised version of SR, namely supervised sparse representation (SSR), is proposed to cooperate with label information and capture the useful structure for classification. Furthermore, we also propose a novel framework of dimension reduction method named discriminative sparse preserving embedding (DSPE). By employing the proposed SSR, DSPE learns a discriminative information from the training set. Moreover, DSPE finds the low dimensional subspace that reduces the within class distances while keeping the between class distances. We evaluate the proposed method on several conditions, including variation of illumination and expression. By comparing with the related methods, the experimental results clearly and consistently show that the proposed framework achieves higher recognition accuracy on benchmark face databases.

Chapter 4

Modular Weighted Global Sparse Representation for Robust Face Recognition

To deal with the large occlusion in the query image, modular SRC divides the image into many modules and processes each module individually. The final decision is based on the voting of all modules. Although modular SRC achieves some impressive results, it is less reliable because of ignoring the correlation between modules by searching the local optimums. Moreover, modular SRC does not differentiate the clean and corrupted modules. To tackle these problems, we proposed a modular weighted global sparse representation (WGSR) framework. The proposed WGSR learns the weight of module based on the modular residual and sparsity jointly, where the corrupted or unreliable module is discarded by setting weight to 0. A reconstructed image from the modules weighted by their reliability is formed for the robust recognition. The proposed framework advances both the modular and global sparse representation approaches, especially in dealing with disguise, large illumination variations and expression changes.

4.1 Introduction

Recently, sparse representation based classification (SRC) [67], which represents the query face image as a sparse linear combination of training samples of all classes, has attracted considerable attention. By replacing the non-convex ℓ_0 -norm with the convex ℓ_1 -norm, SRC achieves impressive recognition results. Different from SRC, linear regression classification (LRC) [71] considers the face recognition task as a linear combination of training samples of a single subject.

Although these methods perform well under the controlled environment, they cannot well handle contiguously occluded face images in the practical scenarios, such as disguise shown in Fig. 4.1a. Therefore, modular approach [72] was applied in both SRC and LRC. They partition the facial image into a number of modules. Each module is processed separately. The final decision is made by fusing the classification result of each module. The modular SRC [67] uses majority voting for classification. The disadvantage of majority voting is that it treats occluded and clean modules equally. In modular LRC [71], authors classify query face image by labeling it to the subject with the minimum class representation error in all modules. However, the limited discriminating power of a single module tends to cause misclassification. The commonality of the above methods is that they essentially treat each module independently. Hence, they lose the correlation information between modules which is very important for classification.

These observations motivate us to explore a more robust approach for face recognition. Firstly, classification should only rely on the clean modules. Secondly, the correlation information between the clean modules should be utilized in the recognition process. Based on these two criteria, we propose a modular weighted global sparse representation (WGSR) method. It is composed of two stages. First, image is divided into modules and each module is processed separately. The modular sparsity and residual are jointly employed to determine the reliability

Figure 4.1: The first row shows normal images. The second row shows image with variations caused by disguise, expression and large illumination changes.

of each module. Then, the modular reliability is used to weight the module for the reconstruction of a global feature vector. Classification is performed on the reconstructed global feature vector.

Besides disguise, the proposed method can also handle large illumination variations and expression changes. While caused by different sources, these variations share some common properties. First, they all have some pixels largely different from those of normal images. Moreover, the distinct pixels are contiguous that form one or a few local areas. As shown in Fig. 4.1, regions in the green boxes are very different from normal images in the first row. The distinct area provides little discriminating or even misleading information for classification. In this work, we try to solve these three typical problems under a common framework that tackles the contiguous occlusion.

4.2 Proposed Modular Weighted Global Sparse Representation

Given a set of n training images from c different subjects, $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n]$, where $\mathbf{a}_j$ is the feature vector representing the j th image. Each training image is partitioned into eight modules as shown in Fig. 4.2, which is the same face image partition choice as [67, 71]. Correspondingly, $\mathbf{A}$ is partitioned as $\mathbf{A} = [\mathbf{A}^1; \dots; \mathbf{A}^l; \dots; \mathbf{A}^8]$ and $\mathbf{A}^l \in \mathbb{R}^{d \times n}$ with $d < n$. Obviously, the j th column of $\mathbf{A}^l$ is the feature vector for the l th module of the j th image. Similarly, a query image $\mathbf{y}$ is partitioned into $\mathbf{y}^1, \dots, \mathbf{y}^l, \dots, \mathbf{y}^8$.

4.2.1 Determination of the Modular Weights

For each module $\mathbf{y}^l$, we treat it as a linear combination of the dictionary $\mathbf{A}^l$ by

$$\mathbf{y}^l = \mathbf{A}^l \mathbf{x}^l. \quad (4.1)$$

As this is an underdetermined linear system ($d < n$), there are infinitely many solutions. To get an unique and stable solution, a constraint is needed. The most popular constraint is ℓ_2 -norm minimization for its simplicity. However, it generates a dense solution that is not informative for recognizing the query image. To achieve a sparse solution, in [67], ℓ_1 -norm minimization is employed.

Applying ℓ_1 -norm minimization, we get the following optimization problems

$$\mathbf{x}^l = \arg \min_{\mathbf{x}} \|\mathbf{A}^l \mathbf{x} - \mathbf{y}^l\|_2^2 + \lambda \|\mathbf{x}\|_1, \text{ for } 1 \leq l \leq 8. \quad (4.2)$$

Figure 4.2: (a) A query face image with scarf. The first row of (b)-(i) shows the eight modules of query image. The second row of (b)-(i) shows the reconstructed image with ℓ_1 -norm minimization. The third row of (b)-(i) shows the estimated modular class coefficients, and the numeric values of modular sparsity and residual. The red coefficients (circle) correspond to correct training subject.

The sparsity should be only measured between classes [141]. Thus, we propose to add up the absolute values of coefficients belonging to the same class to form the modular class coefficient b_i^l for the l th module of the i th class

$$b_i^l = \|\delta_i(\mathbf{x}^l)\|_1, \quad (4.3)$$

where $\delta_i(\mathbf{x}^l)$ is the characteristic function which selects the elements associated with the i th class in $\mathbf{x}^l$.

Concatenate all c modular class coefficients of the l th module to generate the modular class vector $\mathbf{b}^l$ as

$$\mathbf{b}^l = [b_1^l, b_2^l, \dots, b_c^l]. \quad (4.4)$$

We propose the following function for measuring the modular sparsity of the l th module:

$$s^l = \frac{\sqrt{c}}{\sqrt{c}-1} \left(\frac{\|\mathbf{b}^l\|_2}{\|\mathbf{b}^l\|_1} - \frac{1}{\sqrt{c}} \right). \quad (4.5)$$

Here, $\|\mathbf{b}^l\|_2/\|\mathbf{b}^l\|_1$ is a sparsity measurement discussed in [142]. When $\mathbf{b}^l$ has only a single nonzero coefficient, s^l reaches the maximum value 1. When all coefficients of $\mathbf{b}^l$ are the same nonzero value, s^l reaches the minimum value 0.

The modular residual can be measured simply by the ℓ_2 -norm as

$$r^l = \|\mathbf{A}^l \mathbf{x}^l - \mathbf{y}^l\|_2. \quad (4.6)$$

After generating the above measurements, we explore the relation of the occlusion with the modular sparsity and residual. Clean query modules can be accurately represented by only the training samples from the same class. Therefore, they produce large value of the sparsity and

small value of the residual. This is verified by the numeric values of s^1 - s^4 and r^1 - r^4 as shown in Fig. 4.2b-e. The situations of the occluded modules are more complicated. An occluded module could be far away from every subspace spanned by training samples of each class but near or even within the subspace spanned by all training samples of all classes. This results in a small value of the sparsity s^l but also a small value of residual. This is verified by the numeric values of s^7 , s^8 and r^7 , r^8 as shown in Fig. 4.2h and i. While this kind of occluded modules can have even smaller residuals than the clean ones, they can be differentiated by the small values of the sparsity. However, the sparsity alone cannot differentiate all kinds of occluded modules. If an occluded module is significantly nearer to the subspace of one class than the others, the optimization (4.2) will produce, though a large residual, a high value of the sparsity. This is verified by the numeric values of s^5 , s^6 and r^5 , r^6 as shown in Fig. 4.2f and g. Therefore, to differentiate occluded modules from the clean ones, both the modular sparsity and residual should be employed.

A simple way to tackle the occlusion problem is to exclude the occluded modules from the feature vector for the classification. This can be done by multiplying all modules with a weighting function w^l as

$$w^l = \begin{cases} 0, & \text{if } (s^l \leq s_t) \vee (r^l \geq r_t) \\ 1, & \text{if } (s^l > s_t) \wedge (r^l < r_t). \end{cases} \quad (4.7)$$

Obviously, we can compute the weighting function simply by $w^l = w_s^l w_r^l$, where w_s^l and w_r^l are two step functions, respectively switching from 0 to 1 at s_t and from 1 to 0 at r_t .

However, it is difficult to find the optimal hard thresholds s_t and r_t , which could be different from different databases or applications. In addition, it is not difficult to understand that the final classification performance will be very sensitive to the two hard thresholds. Therefore, we relax each hard threshold to two easily found safe thresholds s_1, s_2 and r_1, r_2 so that $(s^l \geq$

$s2) \wedge (r^l \leq r1)$ guarantees a clean module and $(s^l \leq s1) \vee (r^l \geq r2)$ ensures a significantly occluded module. We simply weight modules between these two extreme cases linearly to s^l and r^l . Thus, the weighting function is proposed as

$$w^l = w_s^l w_r^l, \quad (4.8)$$

where

$$w_s^l = \begin{cases} 0, & s^l \leq s1 \\ (s^l - s1)/(s2 - s1), & s1 < s^l < s2 \\ 1, & s^l \geq s2 \end{cases} \quad (4.9)$$

$$w_r^l = \begin{cases} 1, & r^l \leq r1 \\ (r2 - r^l)/(r2 - r1), & r1 < r^l < r2 \\ 0, & r^l \geq r2 \end{cases} \quad (4.10)$$

Fig. 4.3 shows the two weighting values w_s^l and w_r^l as the functions of the sparsity s^l and residual r^l , respectively.

4.2.2 Weighted Global Sparse Representation

The modular representation based methods make a final decision based on the classification result of each module. The loss of the correlation information between modules may cause misclassification. We propose to combine the weighted modules to reconstruct a global feature vector.

Each modular training matrix A^l and modular query vector y^l are weighted by corresponding weighting value to generate a modular weighted training matrix and a modular weighted

Figure 4.3: (a) Weighting function of modular sparsity; (b) Weighting function of modular residual.

query vector:

$$\mathbf{A}_w^l = w^l \mathbf{A}^l, \quad (4.11)$$

$$\mathbf{y}_w^l = w^l \mathbf{y}^l. \quad (4.12)$$

Concatenate all modular weighted training matrices to produce a weighted global training matrix as

$$\mathbf{A}_r = [\mathbf{A}_w^1; \mathbf{A}_w^2; \dots; \mathbf{A}_w^8]. \quad (4.13)$$

Concatenate all weighted modular query vectors to produce a weighted global query vector as

$$\mathbf{y}_r = [\mathbf{y}_w^1; \mathbf{y}_w^2; \dots; \mathbf{y}_w^8]. \quad (4.14)$$

We consider that $\mathbf{y}_r$ is a linear combination of $\mathbf{A}_r$ as

$$\mathbf{y}_r = \mathbf{A}_r \mathbf{x}_r. \quad (4.15)$$

To capture the subject label information from the weighted global matrix and weighted global query vector, we solve the following global ℓ_1 -norm minimization problem

$$\mathbf{x}_r = \arg \min_{\mathbf{x}} \|\mathbf{A}_r \mathbf{x} - \mathbf{y}_r\|_2^2 + \lambda \|\mathbf{x}\|_1. \quad (4.16)$$

It is not difficult to see that $\mathbf{x}_r$ is affected by the modular weighting value w^l . If the heavy weighted module is not well represented, the error will be large. To minimize the global reconstruction error, $\mathbf{x}_r$ tends to reconstruct the modules with larger weights more accurately. Therefore, the reliable modules are heavily weighted. The unreliable modules are weighted by small value down to zero to reduce their impacts on the classification result.

4.2.3 Classification Procedure

Given a weighted global query vector $\mathbf{y}_r$, a weighted global training matrix $\mathbf{A}_r$ and the representation coefficient vector $\mathbf{x}_r$, we allocate the query image to the subject i^* that has the global minimum residual

$$i^* = \arg \min_i \|\mathbf{A}_r \delta_i(\mathbf{x}_r) - \mathbf{y}_r\|_2. \quad (4.17)$$

The following Algorithm 2 summarizes the proposed recognition framework.

Compared with the global LRC and SRC approaches, the proposed WGSR framework attenuates the problems of the query images with corrupted, occluded or largely varied modules

Algorithm 2: Modular Weighted Global Sparse Representation (WGSR)

Input: A set of training images partitioned into eight modules $\mathbf{A}^1, \mathbf{A}^2, \dots, \mathbf{A}^8$ with unit ℓ_2 -norm column. A query image partitioned into 8 modules $\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^8$ with unit ℓ_2 -norm column.

- 1 Solve the modular ℓ_1 -norm minimization problems as equation (4.2).
- 2 Compute the modular sparsity s^l and residual r^l by equations (4.5) and (4.6).
- 3 Compute the modular weights w^l by equations (4.8), (4.9) and (4.10). If all modular weighting values equal 0, set w^{l^*} as 1 for $l^* = \arg \min_l \{\frac{r^l}{s^l}\}$.
- 4 Reconstruct the weighted global training matrix $\mathbf{A}_r$ and weighted global query vector $\mathbf{y}_r$ by equations (4.11), (4.12), (4.13) and (4.14).
- 5 Normalize each reconstructed image to unit ℓ_2 -norm column. Solve the global ℓ_1 -norm minimization problem as equation (4.16).

Output: identify($\mathbf{y}$) = $\arg \min_i \|\mathbf{A}_r \delta_i(\mathbf{x}_r) - \mathbf{y}_r\|_2$.

that may mislead the representation and classification. Compared with the modular based LRC and SRC approaches, the proposed WGSR framework solves the problems of the limited discriminating information of each individual module by seeking the global optimization instead of multiple local optimizations in the modular based LRC and SRC approaches.

4.3 Experiments

We evaluate the performances of the proposed method on the same databases as in [67]: Extended Yale B [143] and AR database [139]. These two face databases focus on frontal faces with illumination variations, disguises and expression changes. In the experiments, we compare the proposed method with the related methods: global SRC [67], modular SRC [67], global LRC [71], modular LRC [71] and relaxed collaborative representation (RCR) [144]. Similar to the modular methods, RCR partitions images into several blocks. While allowing each query module to be flexibly represented, RCR introduces a novel regularization term to enforce the representation vectors having a small variance, and distinguishes the distinctiveness of different modules by adaptive weighting. However, as it optimizes the coefficients and weights iterative-

ly, RCR may converge to a local optimum.

For the proposed method, suitable parameters are studied by cross-validation on the unused 26 subjects of AR database (see the subsection B). $\lambda = 0.05$, $r1 = 0.05$, $r2 = 0.2$, $s1 = 0.2$ and $s2 = 0.8$ are chosen and fixed in all experiments.

4.3.1 Extended Yale B Database

There are 2,414 frontal-face images of 38 subjects with size 168×192 in the cropped Extended Yale B database, which are captured under 64 different lighting conditions. The same as in [67], we randomly select half of the images of each subject for training and the rest are used for testing. To test the effect of the image size on the recognition accuracy, the images are down-sampled to the sizes of $10 \times 12 = 120$, $16 \times 20 = 320$ and $20 \times 24 = 480$. Table 4.1 shows the average face recognition rates over ten runs for different image sizes across various methods.

Table 4.1: Recognition Rate on the Extended Yale B Database.

Image size	10×12	16×20	20×24
global SRC	95.1%	97.1%	97.7%
modular SRC	67.3%	97.0%	98.4%
global LRC	95.2%	96.3%	96.6%
modular LRC	6.2%	73.4%	87.5%
RCR	73.0%	85.7%	94.0%
WGSR	93.7%	98.9%	99%

For the small image size of 10×12 , the performances of the modular approaches are much worse than the global methods due to the insufficient features (dimensionality of 15) for each module. Nonetheless, the recognition rate of the WGSR approach, though utilizing modular information, is much better than the modular methods and close to the global methods. The increase of the image size to 16×20 significantly improves the performances of the modular approaches. However, the advantages of the modular approaches do not yet fully compensate

their disadvantages. The proposed WGSR approach outperforms both the modular and global methods. For the large image size of 20×24 , the performances of the modular approaches are further improved and the modular SRC outperforms its global counterpart. Still, the proposed WGSR framework achieves the best recognition accuracy.

To evaluate the performances of various methods in dealing with different number of training samples per subject, we conduct another experiment in Extended Yale B database. We randomly select t from 64 illuminations as training data and the others are used for testing. We repeat this procedure 10 times for each t (6, 5, 4, 3). As the training number per subject is limited, we use the image with higher dimensionality (i.e., $42 \times 48 = 2016$) to enhance the performances of all methods. Fig. 4.4 details the averaged results of various methods with different training samples t .

As shown in Fig. 4.4, modular LRC underperforms its global version again. However, modular SRC performs much better than global SRC. In this scenario, RCR has similar results to global SRC. The proposed WGSR consistently achieves the best in term of recognition rates. Its highest gain over modular SRC reaches about 5% as t is equal to 3.

4.3.2 AR Database

The AR database consists of over 4,000 frontal-face images from 126 subjects. For each subject, 26 pictures were taken in two separate sessions. In the experiment, same as in [67], 50 male subjects and 50 female subjects are selected.

In the first set of experiment, 799 unoccluded images (about 8 per subject) under varying facial expression as training samples are used. All images are with size of $60 \times 80 = 4800$. The same as in [67], the 4800 dimensional raw pixels are directly used as input without down-sampling. We consider two separate disguise test sets of 200 images. The first disguise set

Figure 4.4: Recognition rate on Extended Yale B versus number of training sample per subject.

contains images of the subjects wearing sunglasses, which occludes roughly 20% of the image. The second disguise set contains images of the subjects wearing a scarf, which occludes roughly 40% of image.

In the second set of experiment, we test the face recognition on AR database with lower resolution. 200 images of neutral expression (two per subject) are chosen for training. 600 images with facial expressions (smile, anger and scream) are selected for testing. All images are down-sampled to the feature vector of the dimensionality of $20 \times 28 = 560$.

Table 4.2 lists the recognition performances. For the two disguise scenarios, as the query samples are severely occluded, the recognition rates of the global methods are much lower than

Table 4.2: Recognition Rate on the AR Database.

Scenarios	Sunglasses	Scarf	Expression
global SRC	87.0%	59.5%	84.3%
modular SRC	97.5%	93.5%	83.8%
global LRC	65.5%	12.5%	85.2%
modular LRC	93.5%	73.5%	82.3%
RCR	98.5%	96.5%	84.3%
WGSR	100%	97.5%	92.4%

those of the modular approaches. The proposed WGSR achieves a gain of 1.5% (1%) over the second best method RCR in the scenario of Sunglasses (Scarf). In the expression changes scenario, however, the modular approaches are inferior to the global ones. One reason could be that the expression changes causes small amount of variation but they spread over most modules. As a result, the advantages of the modular approaches do not fully compensate their problems. In expression scenario, the proposed WGSR framework significantly outperforms both the modular and global approaches.

Fig. 4.5 shows some examples of the query images, weight images and weighted global query images of the proposed WGSR. In Fig. 4.5a and b, it can be seen that the modules occluded by scarf are with small weight (i.e., black) and the clean modules are large weight (i.e., white). Therefore, WGSR discards the lower 4 modules and retains the upper 4 modules as shown in Fig. 4.5c. When the query image is occluded by sunglasses, a similar result is achieved, which is shown in Fig. 4.5d, e and f.

4.3.3 Multi-PIE

The CMU Multi-PIE database contains face images captured in 4 sessions with varying illuminations, expressions and poses. In this experiment, similar to Section 4.3.2, we test the

Figure 4.5: The query images and their corresponding weight images and weighted global query images. (a) Query image with scarf; (b) weight image of (a); (c) the weighted global query image of (a); (d) query image with sunglasses; (e) weight image of (d); (f) the weighted global query image of (d). In (b) and (e), black and white pixels denote 0 and 1 weight, respectively.

performances of different methods on face images with expression variation. The face images with only expression variation from the first 105 subjects that appear in all first 3 sessions are used. Images are cropped based on the eye locations provided by [145] and down-sampled to 60×40 pixels. The face images with neutral expression from all 3 sessions are used as training data, the rest are used for testing. That is, for each subject, there are three training images and five testing images (2 smile, 1 surprise, 1 squint and 1 disgust). Fig. 4.6 shows some faces with different expressions in CMU Multi-PIE.

Table 4.3 details the recognition rates of different expressions and overall results. The overall results of this experiment are consistent with those of Section 4.3.2, where the global methods outperform the modular ones. The overall recognition rates of RCR are slightly worse than those of global SRC and global LRC. It can be seen that the proposed WGSR achieves the best overall result, which obtains a gain of 3.8% over the second best one. In details, the proposed methods perform as the best in all except for surprise expression, where it is the second best.

Figure 4.6: Faces from Multi-PIE database with different expressions.

Table 4.3: Recognition Rate on the Multi-PIE Database.

Scenarios	Smile	Surprise	Squint	Disgust	Overall
global SRC	96.7%	81.0%	91.4%	83.8%	89.9%
modular SRC	90%	82.9%	91.4%	62.9%	83.4%
global LRC	98.1%	81.9%	87.6%	81%	89.3%
modular LRC	82.4%	80%	75.2%	59.1%	75.8%
RCR	91.4%	86.7%	90.5%	79.2%	87.2%
WGSR	98.6%	85.7%	97.1%	88.6%	93.7%

4.4 Summary

In this chapter, we proposed a novel robust face recognition framework that employs the modular approach to reconstruct a weighted global feature vector for sparse representation. Classification is performed on the reconstructed global representation of the face image. This framework attenuates problems of both the modular and global representation methods. We evaluate the proposed WGSR on several conditions, including variation of illumination, expression changes and two different disguises. The experimental results clearly and consistently show that the proposed WGSR is much more robust than both the modular and global representation methods.

Chapter 5

Class-wise Sparse and Collaborative Patch Representation for Robust Face Recognition

Although Chapter 3 introduced a supervised sparse representation (SSR), it is only applicable when the label of the reconstructed sample is known. As a result, it limits to training data and cannot directly apply to classification. To alleviate the problem of unsupervised optimization of SRC, we proposed a class-wise sparse representation (CSR) by employing the class label information. The proposed CSR seeks an optimum representation of the query image such that it minimizes the reconstruction error and maximizes class-wise sparsity simultaneously. To tackle the problem of heavily contaminated training data, we further propose a collaborative patch framework, together with CSR, named CSR-CP. Different from the patch based methods, which optimize each patch independently, the proposed CSR-CP reconstructs the whole query sample by sub-images of all subjects across all pathes within one framework and searches a patch-wise and class-wise sparse representation. As a result, a more reliable and discriminative result can be attained.

5.1 Introduction

If images originate from distinct subjects are less correlated than those from the same subject, a sparse representation is expected with the significant representation coefficients focusing on the correct subject. However, the face subspaces spanned by different subjects do have some correlation due to the common facial components (e.g., eyes, nose, and mouth). Moreover, face images of similar poses, expressions or illumination conditions from different subjects may show higher correlation than those from the same subject under different conditions. Therefore, the unsupervised ℓ_1 -norm minimization of the sample coefficients in SRC may result in the representation of a query face by samples from many different subjects, which may cause misclassification. To alleviate this problem, Group Lasso (GL) [101] is proposed to seek an image representation with minimized $\ell_{2,1}$ -norm of the sample coefficients. However, the $\ell_{2,1}$ -norm not only minimizes the number of the selected subjects, but also minimizes the ℓ_2 -norm of coefficients within each class. The later may prevent the attained result from the desired solution as the optimal representation of a query image by training samples of the correct subject may not necessarily be dense.

The first part of this work aims at alleviating the problems of SRC and GL. Towards this end, we propose a new algorithm, named class-wise sparse representation (CSR). Instead of minimizing the number of training samples used in SRC, it minimizes the number of training classes in representing the query image. Different from the GL, the proposed CSR has more emphasis on the sparsity between the classes and hence more likely selects the correct class. In implementation, the desired solution can be obtained by optimizing the representation error and the class-wise sparsity simultaneously, which can be formulated as a convex problem solved via Augmented Lagrange Multiplier scheme [146, 147].

The holistic approaches have been demonstrated being effective for uncorrupted training

and testing data. However, they are extremely vulnerable to the uncontrolled data (e.g., extreme shadows and disguise) [148–150]. Local approaches, such as keypoint based [151–153], component based [154–156] and patch based [67, 71, 72, 157] methods, are proven to be more robust to the disguised or occluded images. However, both keypoint and component based methods have huge computational burden and their good performances heavily depend on the reliable detection of the keypoints or components, which is not an easy task. Patch based approaches that simply partition the image into predefined patches is often preferred for its usability. Conventional patch based approaches apply the classifier to each patch separately and fuse the patch classification results for final decision. Many fusion methods have been studied, such as product rule, sum rule, max rule, median rule, and weighted sum rule. In [71], LRC is applied to each patch and the label of the query image follows the result of patch that has the minimum representation error. Differently, SRC by patch [67] applies the majority voting to the classification results of individual patches. It treats all patches equally regardless whether they are corrupted or not. To alleviate this problem, modular weighted global sparse representation (WGSR) [157] discards the corrupted patches based on the patch residual and sparsity, and uses the remaining patches for classification by weighting patches with a nonlinear function. However, the optimization of the weighting function remains an open issue.

As each patch is processed separately, patch based approach is more flexible than holistic one. Without the bounding between patches, the corrupted patches will not affect the representation and classification of the uncorrupted ones. Unfortunately, the patch scheme largely reduces the discriminative information used in the representation and classification. The classification result based on the single patch is much less reliable than that from the whole image. Fusing these less reliable results does not guarantee a correct classification.

To exploit the flexible representation of patches and alleviate the unreliability problem caused by the separated patch optimization, a collaborative patch (CP) representation scheme,

together with the proposed CSR, named CSR-CP, is proposed in the second part of this chapter. Different from the holistic approaches that restrict the representation coefficient being the same for all pixels of a training image, each patch in the proposed scheme has its own representation coefficient. Different from the aforementioned patch based approaches that optimize each patch representation separately, the CP scheme puts all patches of an image into a group and optimizes a group-wise sparse representation of the whole query image. By seeking the group-wise sparsity over the coefficients of all patches, it harnesses the relationship among patches of the same image to obtain a reliable and discriminative result. The proposed CP scheme can be realized by extending the proposed CSR algorithm to the group of patches. Comprehensive experimental results are presented in this chapter to verify the effectiveness of the proposed CSR-CP approach on several benchmark databases. They show that it greatly enhances the classification robustness in dealing with uncontrolled and insufficient training data.

5.2 Class-wise Sparse Representation

An image of m pixels is arranged in a column vector. Let $\mathbf{A}_i = [\mathbf{a}_{i,1}, \dots, \mathbf{a}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$ be a stack of n_i training samples with dimensionality of m from the i th subject. $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_C] \in \mathbb{R}^{m \times n}$ stacks training samples of all C subjects and $n = \sum_{i=1}^C n_i$. A query image $\mathbf{y} \in \mathbb{R}^m$ can be represented by training samples of all classes $\mathbf{A}$ as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \quad (5.1)$$

where $\mathbf{x} \in \mathbb{R}^n$ is a representation coefficient vector associated with $\mathbf{A}$, and $\mathbf{e}$ is the representation error. As it is in general assumed that samples of a specific subject lie in a linear subspace, a query image is expected to be well represented as a linear combination of the training samples of the same class. Thus, we can find a sparse solution to the representation coefficients $\mathbf{x}$. By

measuring the sparsity of the coefficient vector with ℓ_1 -norm, the sparse optimization problem is formulated as

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1, \quad (5.2)$$

where $\|\cdot\|_1$ denotes the ℓ_1 -norm that is absolute sum of a vector and λ is a parameter for compromise between the reconstruction error and sample sparsity. Many algorithms are proposed to solve problem (5.2) efficiently [82, 86, 88]. With the assumption that subspaces of distinct classes are independent to each other, (5.2) achieves a discriminative representation where significant nonzero coefficients are only associated to the correct subject [53, 118]. Therefore, sparse representation (5.2) is applied to classification tasks, such as face recognition, and named sparse representation based classification (SRC) scheme [67].

However, there are two problems of SRC for face recognition: First, the significant nonzero coefficients, though sparse, could be associated with many subjects. Due to the common facial components, the assumption of independence of face subspaces cannot be always guaranteed, which leads to significant nonzero coefficients spreading into many distinct subjects. Second, to achieve a sparse coefficient, SRC tends to randomly select a single representative from the highly correlated training samples [158, 159]. The facial images have not only the identity information but also much other information, such as illumination and expression. Thus, the correlation between the samples of some different subjects with similar variations could be higher than those of the same subject with different variations. In this case, SRC is known to have stability problems [159]. Typically, given two highly correlated samples, SRC will randomly select one of the two [159]. The randomness of SRC may lead to an unreliable result.

Although these two problems seem unrelated, both of them can be resolved by using the class labels of training data during sparse optimization. With this discriminative information, the significant representation coefficients can be forced to associate with few subjects. Selecting classes instead of individual samples can alleviate the problem of random selection of highly

correlated data. Therefore, Group Lasso (GL) [101] is proposed to conquer the problems of SRC by replacing the ℓ_1 -norm with the so-called $\ell_{2,1}$ -norm as following

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_{i=1}^C \|\mathbf{x}_i\|_2, \quad (5.3)$$

where $\mathbf{x}_i \in \mathcal{R}^{n_i}$ is the representation coefficient vector of the i th class and $\mathbf{x} = [\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_C]$. The second term of (5.3) is $\ell_{2,1}$ -norm [160] that is widely used for group sparsity measurement.

The $\ell_{2,1}$ -norm is in fact the combination of a ℓ_2 -norm within class and a ℓ_1 -norm across classes. It is not difficult to understand the role of ℓ_1 -norm minimization over classes, which enforces a group sparse representation. However, the minimization of ℓ_2 -norm suppresses significant sparse coefficients and promotes a dense representation within the class. Take Fig. 5.1 as an example, where four training samples (i.e., $\mathbf{a}_{1,1}$, $\mathbf{a}_{1,2}$, $\mathbf{a}_{2,1}$, $\mathbf{a}_{2,2}$) of two subjects (triangle and diamond) and a query sample $\mathbf{y}$ (circle) are shown. The query sample can be well represented by the samples of the triangle subject, i.e., $\mathbf{y} = 0.97 \cdot \mathbf{a}_{1,1} + 0.01 \cdot \mathbf{a}_{1,2}$. It can also be well represented by a combination of samples of the diamond subject as $\mathbf{y} = 0.5 \cdot \mathbf{a}_{2,1} + 0.5 \cdot \mathbf{a}_{2,2}$. Although the ℓ_1 -norm of the former is smaller than the later, its ℓ_2 -norm (0.970) is much larger than the later (0.707). Therefore, the ℓ_2 -norm minimization of GL results in the death of the single large coefficient and the survival of the two small weights. The former one, however, is more likely to be the correct subject because the face image variation is more likely sparsely distributed rather than a dense Gaussian distribution. Therefore, the ℓ_2 -norm minimization in the second term of (5.3) adversely affects the correct class selection during the minimization process of GL.

To tackle the problems of SRC and GL, we propose the following optimization problem:

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{z}} \quad & \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{z}\|_0 \text{ s.t. } z_i = \|\mathbf{x}_i\|_2, \\ \text{where } \mathbf{z} = & [z_1, z_2, \dots, z_C]^T. \end{aligned} \quad (5.4)$$

Figure 5.1: An example of two different representations: a query image y (circle) can be sparsely well represented by samples $a_{1,1}$ and $a_{1,2}$ of one class (triangle); it can also be densely well represented by samples $a_{2,1}$ and $a_{2,2}$ of another class (diamond).

Here the first term is the reconstruction error, the second term is a class-wise sparsity measurement, and λ is a parameter balancing the effects of these two terms.

The minimization of ℓ_0 -norm over $\mathbf{z}$ in (5.4) is a non-convex problem, which can only be solved by the exhaust searching [161]. Fortunately, recent studies in compress sampling [39, 74, 162] show that, the sparse vector can be approximately recovered by replacing the non-convex ℓ_0 -norm with the convex ℓ_1 -norm if the solution is sparse enough. To be robust to the gross images, which is ubiquitously contaminated by sparse noises with arbitrary magnitudes [117] and therefore cannot be well characterized by the ℓ_2 -norm, ℓ_1 -norm is used to measure the representation error. As a result, we use ℓ_1 -norm minimization for both sparse reconstruction error and class-wise sparsity. Therefore, we propose a class-wise sparse representation (CSR) method as

$$\min_{\mathbf{x}, \mathbf{z}} \frac{1}{2} \|\mathbf{y} - \mathbf{Ax}\|_1 + \lambda \|\mathbf{z}\|_1 \text{ s.t. } z_i = \|\mathbf{x}_i\|_2, \\ \text{where } \mathbf{z} = [z_1, z_2, \dots, z_C]^T. \quad (5.5)$$

The proposed CSR introduces a new variable $\mathbf{z}$ into the optimization process. The minimization of the number of nonzero elements of the C dimensional vector $\mathbf{z}$ directly coincides with the aim of finding the most class-wise sparse representation of C classes. In contrast, GL searches a group sparse representation $\mathbf{x}$ in $\mathbb{R}^n$ space. Its minimization of the ℓ_2 -norm of $\mathbf{x}_i$ may lead to an undesirable result as the optimal representation of a query image by training samples of the correct subject may not necessarily be dense. In addition, it is troublesome to minimize the ℓ_2 -norm of the representation error if there are sparse but large corruptions in images. By characterizing the representation error with the ℓ_1 -norm in the proposed CSR, the adversarial impact of large noise is suppressed.

Fig. 5.2 visualizes a comparison of coefficients of SRC, GL, and the proposed CSR. Fig. 5.2a shows that SRC achieves the most sample sparse coefficients with only 17 nonzero elements. However, they spread into 9 different subjects as it does not use the class label information of training data. The most significant coefficient is associated with the wrong subject that leads to misclassification in this example. Fig. 5.2b shows coefficients of GL, in which 7 subjects have nonzero coefficients. The ℓ_2 -norm leads to a dense representation within the class. The correct class fails to win the most significant representation coefficient, which results in the wrong decision. For the proposed CSR, it achieves the most class-wise sparse coefficients (5 classes) among the three methods. Furthermore, it makes the correct classification by assigning the correct subject with the largest coefficient, which stand out against the others. Fig. 5.2 shows that the proposed CSR achieves the most discriminative representation among the three.

Figure 5.2: Comparison of representation coefficients of SRC (a), GL (b), and the proposed CSR (c) on AR database of 20 persons with 5 undisguised training images per person. The x-axis is $5(k-1) + j$ where k and j are respectively the indices of subjects and training images of a subject. Coefficients of the correct subject are in red lines (circles) and others are in blue lines (dots).

5.2.1 Optimization

To solve problem (5.5), we first convert it to the following equivalent optimization problem by introducing two auxiliary variables $\mathbf{e} \in \mathbb{R}^m$ and $\mathbf{u} \in \mathbb{R}^n$:

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{z}, \mathbf{e}, \mathbf{u}} \quad & \frac{1}{2} \|\mathbf{e}\|_1 + \lambda \|\mathbf{z}\|_1 \\ \text{s.t.} \quad & \mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x}, \mathbf{x} = \mathbf{u}, z_i = \|\mathbf{u}_i\|_2, \end{aligned} \quad (5.6)$$

where $\mathbf{u}_i \in \mathbb{R}^{n_i}$ is a sub-vector of $\mathbf{u}$ with coefficients associated with class i .

For efficiency, we adopt the Augmented Lagrange Multiplier scheme [146, 147]. Consequently, it derives the following unconstrained optimization problem:

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{z}, \mathbf{e}, \mathbf{u}} \quad & \frac{1}{2} \|\mathbf{e}\|_1 + \lambda \|\mathbf{z}\|_1 \\ & + \gamma^T (\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}) + \boldsymbol{\theta}^T (\tilde{\mathbf{u}} - \mathbf{z}) + \boldsymbol{\beta}^T (\mathbf{x} - \mathbf{u}) \\ & + \frac{\mu}{2} (\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_2^2 + \|\tilde{\mathbf{u}} - \mathbf{z}\|_2^2 + \|\mathbf{x} - \mathbf{u}\|_2^2) \end{aligned} \quad (5.7)$$

where $\boldsymbol{\gamma} \in \mathbb{R}^m$, $\boldsymbol{\theta} \in \mathbb{R}^C$ and $\boldsymbol{\beta} \in \mathbb{R}^n$ are the Lagrangian multipliers, $\mu > 0$ is the penalty parameter, and $\tilde{\mathbf{u}} \in \mathbb{R}^C$ denotes the vector $[\|\mathbf{u}_1\|_2, \|\mathbf{u}_2\|_2, \dots, \|\mathbf{u}_C\|_2]^T$.

Instead of optimizing all variables simultaneously, as they are separable, we apply the alternating optimization scheme, which has been widely advocated by sparse and low-rank optimization works [114, 118, 150, 163]. As a result, the original problem is decomposed into several subproblems, which optimize one variable by fixing the others in each step.

Given the current $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{u}$, we optimize $\mathbf{e}$ by solving the following subproblem:

$$\begin{aligned} \min_{\mathbf{e}} & \frac{1}{2} \|\mathbf{e}\|_1 + \gamma^T(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_2^2 \\ & = \frac{1}{2\mu} \|\mathbf{e}\|_1 + \frac{1}{2} \|\mathbf{e} - (\mathbf{y} - \mathbf{A}\mathbf{x} + \gamma/\mu)\|_2^2. \end{aligned} \quad (5.8)$$

Problem (5.8) can be solved via the soft-thresholding operator [79], $S(\mathbf{y} - \mathbf{A}\mathbf{x} + \gamma/\mu, 1/2\mu)$.

Here $S(\cdot, \cdot)$ is defined element-wise as follows:

$$S(\mathbf{t}, \omega)_q = \text{sign}(t_q) \cdot \max\{|t_q| - \omega, 0\}, \text{ for } q = 1, \dots, m. \quad (5.9)$$

We fix the other variables and optimize $\mathbf{x}$ by minimizing the following problem:

$$\begin{aligned} \min_{\mathbf{x}} & \gamma^T(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}) + \beta^T(\mathbf{x} - \mathbf{u}) \\ & + \frac{\mu}{2} (\|\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}\|_2^2 + \|\mathbf{x} - \mathbf{u}\|_2^2) \\ & = \frac{\mu}{2} \mathbf{x}^T (\mathbf{A}^T \mathbf{A} + \mathbf{I}) \mathbf{x} \\ & - \mu ((\mathbf{y} - \mathbf{e} + \gamma/\mu)^T \mathbf{A} + (\mathbf{u} - \beta/\mu)^T) \mathbf{x}. \end{aligned} \quad (5.10)$$

This is a least square problem, which leads to a close-form solution of $\mathbf{x}$ as:

$$\begin{aligned} \mathbf{x} & = (\mathbf{A}^T \mathbf{A} + \mathbf{I})^{-1} \\ & \quad (\mathbf{A}^T (\mathbf{y} - \mathbf{e} + \gamma/\mu) + (\mathbf{u} - \beta/\mu)). \end{aligned} \quad (5.11)$$

With the updated $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{e}$, the optimized $\mathbf{u}$ can be obtained by

$$\begin{aligned} \min_{\mathbf{u}} & \theta^T(\tilde{\mathbf{u}} - \mathbf{z}) + \beta^T(\mathbf{x} - \mathbf{u}) \\ & + \frac{\mu}{2} (\|\tilde{\mathbf{u}} - \mathbf{z}\|_2^2 + \|\mathbf{x} - \mathbf{u}\|_2^2). \end{aligned} \quad (5.12)$$

By simple manipulation, (5.12) can be rewritten as

$$\min_{\mathbf{u}} \sum_{i=1}^C \left[(\theta_i - \mu z_i) \|\mathbf{u}_i\|_2 + \mu \|\mathbf{u}_i - (\mathbf{x}_i + \boldsymbol{\beta}_i/\mu)/2\|_2^2 \right], \quad (5.13)$$

which has a closed form solution by the one-dimensional shrinkage formula [160]:

$$\mathbf{u}_i = \max\left(\|\mathbf{r}_i\|_2 - \frac{(\theta_i - \mu z_i)}{2\mu}, 0\right) \frac{\mathbf{r}_i}{\|\mathbf{r}_i\|_2}, \quad (5.14)$$

where $\mathbf{r}_i = (\mathbf{x}_i + \boldsymbol{\beta}_i/\mu)/2$, for $i = 1, 2, \dots, C$.

We can acquire the optimum $\mathbf{z}$ by solving the following problem:

$$\begin{aligned} \min_{\mathbf{z}} \lambda \|\mathbf{z}\|_1 + \boldsymbol{\theta}(\tilde{\mathbf{u}} - \mathbf{z}) + \frac{\mu}{2} \|\tilde{\mathbf{u}} - \mathbf{z}\|_2^2 \\ = \frac{\lambda}{\mu} \|\mathbf{z}\|_1 + \frac{1}{2} \|\mathbf{z} - (\tilde{\mathbf{u}} + \boldsymbol{\theta}/\mu)\|_2^2. \end{aligned} \quad (5.15)$$

The same as problem (5.8), (5.15) can be solved by the soft-thresholding operator.

The Lagrangian multipliers are updated as

$$\boldsymbol{\gamma} = \boldsymbol{\gamma} + \mu(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{e}), \quad (5.16)$$

$$\boldsymbol{\theta} = \boldsymbol{\theta} + \mu(\tilde{\mathbf{u}} - \mathbf{z}), \quad (5.17)$$

$$\boldsymbol{\beta} = \boldsymbol{\beta} + \mu(\mathbf{x} - \mathbf{u}). \quad (5.18)$$

The steps (5.8), (5.10), (5.13), and (5.15)-(5.18) are repeated until the convergence conditions are attained. Algorithm 3 summarizes the procedures of solving the optimization problem (5.5). GL employs two steps similar to (5.10) and (5.13) of the proposed CSR. However, com-

pared to the proposed CSR, GL lacks two crucial steps (5.8) and (5.15). In (5.8), the proposed CSR attempts to find the sparse noise vector $\mathbf{e}$. It is well known that the l_2 -norm minimization (5.10) is not robust to outliers, especially, when the outliers are significantly different from the original data. By removing the sparse but arbitrary noise via (5.8), the optimized $\mathbf{x}$ in (5.10) will less overfit to a corrupted query sample $\mathbf{y}$. In (5.15), we minimize the l_1 -norm of $\mathbf{z}$ by searching z_i and setting the small nonzero elements of $\tilde{\mathbf{u}}$ to zero. As a result, the proposed CSR achieves more class-wise sparse coefficients compared to GL.

Thanks to the additional steps (5.8) and (5.15), the proposed CSR outperforms GL, which is further verified in the experiments.

Algorithm 3: Class-wise Sparse Representation (CSR)

Input: Matrixes of training sample set $\mathbf{A}$, query image $\mathbf{y}$, parameter λ .

Initialization:

$\mathbf{x} = \mathbf{0}, \mathbf{z} = \mathbf{0}, \mathbf{u} = \mathbf{0}, \gamma = \mathbf{0}, \boldsymbol{\theta} = \mathbf{0}, \boldsymbol{\beta} = \mathbf{0}, \mu = 10^{-3}, \mu_{max} = 10^6, \rho = 1.1, \epsilon = 10^{-3}$.

while not converged do

1. update the $\mathbf{e}$ as Problem (5.8).
2. update the $\mathbf{x}$ as Problem (5.10).
3. update the $\mathbf{u}$ as Problem (5.13).
4. update the $\mathbf{z}$ as Problem (5.15).
5. update the multipliers as (5.16)-(5.18).
6. update the parameter μ by $\mu = \min(\rho\mu, \mu_{max})$.
7. check the convergence conditions: $\|\mathbf{y} - \mathbf{Ax} - \mathbf{e}\|_\infty < \epsilon$ and $\|\tilde{\mathbf{u}} - \mathbf{z}\|_\infty < \epsilon$ and $\|\mathbf{x} - \mathbf{u}\|_\infty < \epsilon$.

end

Output: $\mathbf{x}, \mathbf{e}$.

5.2.2 Classification

Once the coefficient vector $\mathbf{x}$ and representation error $\mathbf{e}$ are obtained, the corrupted query image $\mathbf{y}$ can be recovered as

$$\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = \mathbf{Ax}. \quad (5.19)$$

Now we use training samples of a single class to reconstruct the recovered image $\hat{\mathbf{y}}$ as

$$\hat{\mathbf{y}} = \mathbf{A}\delta_i(\mathbf{x}) + \mathbf{e}_i, \quad (5.20)$$

where $\delta_i(\mathbf{x})$ only retains the coefficients associated with class i and the others are 0, $\mathbf{e}_i$ is reconstruction error by a single class.

The reconstruction by samples of class i will be the closest to $\hat{\mathbf{y}}$ if the query image originates from the same class. Thus, the reconstruction error of the correct class should be the minimum among all. To measure the magnitude of class reconstruction error, ℓ_2 -norm is widely applied. However, the quadratic term of ℓ_2 -norm makes it easily disturbed by a few or even a single large outlier. To tackle this problem, ℓ_1 -norm is a popular choice in field of robust statistics. In this chapter, we propose to measure the magnitude of class error as

$$r_i = \|\mathbf{e}_i\|_1 = \|\mathbf{A}\mathbf{x} - \mathbf{A}\delta_i(\mathbf{x})\|_1 = \|\mathbf{A}\bar{\delta}_i(\mathbf{x})\|_1, \quad (5.21)$$

where $\bar{\delta}_i(\mathbf{x}) \in \mathbb{R}^n$ is a vector generated by setting the elements of $\mathbf{x}$ associated with the i th class to 0.

Finally, the query image is labeled to the subject with the minimum residual

$$i^* = \arg \min_i r_i. \quad (5.22)$$

5.3 Collaborative Patch Representation

SRC [67] has demonstrated its effectiveness in face recognition under some scenarios. However, it is vulnerable to the insufficient training data [148, 149]. In addition, a violation of carefully controlled training samples may result in severe performances degradation. For example, to

Figure 5.3: (a) A query image partitioned into the predefined patches. (b) An example of collaborative patch sparse representation of the whole query image. The coefficients indicated by red lines (circle) are associated with the correct class.

represent an uncorrupted query image, instead of choosing the corrupted training samples of the correct subject, SRC tends to select the uncorrupted samples from other subjects. In this case, the result of SRC is no longer informative.

In order to deal with face corrupted by large disguise or continuous occlusion, in [67], images are partitioned into L patches as shown in Fig. 5.3a. It generates a set of patch dictionaries $\mathbf{A}^l \in \mathbb{R}^{(m/L) \times n}$ by stacking the l th patches of n training images, for $l = 1, 2, \dots, L$. Similarly, a query image is also partitioned into $\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^L \in \mathbb{R}^{m/L}$. For each patch, $\mathbf{y}^l$ is represented by the patch dictionary $\mathbf{A}^l$ as

$$\mathbf{y}^l = \mathbf{A}^l \mathbf{x}^l + \mathbf{e}^l, \quad (5.23)$$

where $\mathbf{x}^l \in \mathbb{R}^n$ and $\mathbf{e}^l \in \mathbb{R}^{m/L}$ are the patch representation coefficients and error, respectively. For each of L patches, SRC (5.2) is applied to (5.23) for classification. The label of query image is decided based on the voting of all patches [67]. In this chapter, we call this method SRC by patch (SRC-P).

SRC-P alleviates some limitations of SRC. By partitioning the image, the corrupted patches will not affect the representation and classification of the uncorrupted ones. Even if all training samples of the correct subject are partially occluded but at different locations, they still can well

represent the query image by patch. For each patch, SRC-P solves a linear system of $(m/L) \times n$, which has only one L th constraints of SRC ($m \times n$) and hence it reconstructs the query image more accurately.

SRC-P searches individual local optimums first (i.e., the result of each patch) and then fuse them for a final decision. The discriminative information of a patch with the dimensionality of m/L is much less than that of the whole image with the dimensionality of m . Based on the reduced discriminative information, the intermediate result of SRC-P is less reliable. The fusion of these less reliable results does not guarantee a correct classification. Moreover, as all patches are processed separately, they are not constrained to select the same classes. Thus, the significant representation coefficients of different patches could be distributed in a number of different training classes with no class winning more patches than the others.

To alleviate this problem, we propose a scheme that gathers all patches into a single joint optimization. Given a training image $\mathbf{a}_{i,k}$, we generate L patch images, $\mathbf{a}_{i,k}^1, \mathbf{a}_{i,k}^2, \dots, \mathbf{a}_{i,k}^L \in \mathbb{R}^m$, where the only nonzero elements of $\mathbf{a}_{i,k}^l$ are the pixels of $\mathbf{a}_{i,k}$ associated with the l th patch. Some examples of $\mathbf{a}_{i,k}^l$ are shown in the left image of Fig 5.3b, in which the patch images of a row are generated by an image. We stacks all L patch images of $\mathbf{a}_{i,k}$ as $\mathbf{A}_{i,k}^P = [\mathbf{a}_{i,k}^1, \mathbf{a}_{i,k}^2, \dots, \mathbf{a}_{i,k}^L]$. $\mathbf{A}_i^P = [\mathbf{A}_{i,1}^P, \mathbf{A}_{i,2}^P, \dots, \mathbf{A}_{i,n_i}^P]$ encloses all patch images of a subject. By concatenating the patch images of all C subjects, a dictionary is produced as $\mathbf{A}^P = [\mathbf{A}_1^P, \mathbf{A}_2^P, \dots, \mathbf{A}_C^P] \in \mathbb{R}^{m \times nL}$. We propose a collaborative patch (CP) representation of query sample $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{A}^P \mathbf{x}^P + \mathbf{e}^P, \quad (5.24)$$

where $\mathbf{x}^P \in \mathbb{R}^{nL}$ and $\mathbf{e}^P \in \mathbb{R}^m$ are the collaborative patch representation coefficients and error, respectively.

It is not difficult to see that if the proposed CSR (5.5) is applied to solve $\mathbf{x}^P$ in (5.24), the

representation coefficients of all patches of a training image are put into a group to compete with others in the sparse optimization. Thus, all patches of an image collaboratively participate in the optimization. Therefore, we propose the CP representation scheme together with the proposed CSR, named CSR-CP, as following optimization problem

$$\begin{aligned} \min_{\mathbf{x}^P, \mathbf{z}^P} \quad & \frac{1}{2} \|\mathbf{y} - \mathbf{A}^P \mathbf{x}^P\|_1 + \lambda \|\mathbf{z}^P\|_1 \text{ s.t. } z_i^P = \|\mathbf{x}_i^P\|_2 \\ \text{where } \mathbf{z}^P = & [z_1^P, z_2^P, \dots, z_C^P]^T, \end{aligned} \quad (5.25)$$

where $\mathbf{x}_i^P \in \mathbb{R}^{n_i L}$ is the coefficient vector associated with $\mathbf{A}_i^P$. All patches in $\mathbf{A}^P$ and $\mathbf{y}$ are normalized to unit ℓ_2 -norm. Different from SRC-P, which considers each patch representation separately, the proposed CSR-CP method bounds all patches of an image and all images of a class into a group. Putting all patches of an image together in the optimization, CSR-CP compensates the loss of the discriminative information caused by partition of the image into patches. Thus, the result of the proposed CSR-CP is more informative than SRC-P for classification.

Once (5.25) is solved by Algorithm 1, a classification procedure similar to (5.21) and (5.22) is applied.

For SRC, all pixels of an image are constrained to share the same coefficient. Thus, SRC either selects all or none pixels of an image. SRC-P attempts to interpolate between these two conditions by partitioning image into patches and processing them separately. As a result, every patch has its own representation coefficient. In this way, the uncorrupted patches of the corrupted image can be freely and independently used to reconstruct those of query image. Therefore, SRC-P shows its effectiveness when the training data is not well controlled, such as with disguise, extreme expression and shadow. Although SRC-P can achieve a more accurate representation, it does not necessarily result in better recognition rates in all cases. In fact, SRC-P leads the representation to an extreme, which does not take the relations between patches of the same image into consideration. Hence, SRC-P largely reduces the discriminative informa-

Figure 5.4: Comparison of SRC (a), SRC-P (b) and the proposed CSR-CP (c) in terms of representation accuracy by single class. For SRC-P, the class residual r_i is the sum of the 8 patch residuals for comparison with SRC and CSR-CP. The red lines (circle) are associated with the correct subject.

tion, which is critical for classification. The proposed CSR-CP makes a compromise between SRC that treats the whole image as an entirety, and SRC-P that allows different representations for different patches. The patch scheme makes the proposed method flexible in dealing with the corrupted and insufficient training data. The proposed CSR-CP uses all patches collaboratively to learn a group-wise sparse representation where patches are united as an entirety in the sparse optimization. The proposed CSR-CP, which inherits some merits of both holistic and patch based approaches, attains a more reliable and robust representation.

To illustrate a problem the proposed approach solves, Fig. 5.4 shows a real example comparing CSR-CP with SRC and SRC-P. An undisguised query image is represented by training samples that contains 100 subjects, each of which has two images, from AR databases. The two training images of the correct subject are both disguised, one wearing sunglasses and the other wearing scarf. SRC-P and the proposed CSR-CP apply 8 patches as shown in Fig. 5.3. The representation accuracy $(1 - r_i / \max(r_i))$ by training samples of a single class of SRC, SRC-P and the proposed CSR-CP are shown in Fig. 5.4a, b and c, respectively. Due to the disguise in training samples of the correct subject, SRC tends to select the uncorrupted samples of many other subjects to represent the query image. As a result, there are 4 class-wise representation accuracies higher than that of the correct subject, which leads to wrong classification. As SRC-P optimizes patches separately, the different patches have different selections of subjects. This can be seen in Fig. 5.4b that many subjects have nonzero class-wise representation accuracies, where 2 subjects have higher class-wise representation accuracies than the correct subject. More specifically, the minimum reconstruction errors of 8 patches are won by 8 different subjects that leads to misclassification. For the proposed CSR-CP method, the correct subject achieves a much higher class-wise representation accuracy than the other subjects and hence it correctly labels the query image.

5.4 Experiments

The proposed approaches, class-wise sparse representation (CSR) and class-wise sparse representation with collaborative patch (CSR-CP), are evaluated in 4 face databases: Extended Yale B [143], CMU Multi-PIE [164], AR [139], and CMU PIE [140]. To show the effectiveness of the proposed CP separately, we also propose to integrate CP scheme in the Group Lasso [101] named GL-CP. All these three proposed algorithms are compared with LRC [71], SRC [67], collaborative representation based classification with l_1 residual (CRC) [165], Group Lasso (GL) [96, 101], LRC by patch (LRC-P) [71], SRC by patch (SRC-P) [67], relaxed collaborative representation (RCR) [144] and WGSF [157]. The parameter λ of the proposed methods are determined by cross-validation on AR dataset and fixed over all experiments of this paper to 0.25 and 0.1 for CSR (5.5) and CSR-CP (5.25), respectively, though better performances may be achieved if they are fine tuned to fit each specific experiment. For the competing methods, their parameters are tuned to achieve the best performances on AR dataset. The gray level of image is used as the feature. In all but Experiment 7, the number of patch is fixed to 8 as shown in Fig. 5.3, which is the same as [67, 71, 157].

The cropped Extended Yale B database includes 38 subjects, each of which has 64 frontal face images captured under 64 different illuminations. These 64 samples are divided into 5 subsets depending on the angle between the directions of light and face. Fig. 5.5 shows some samples with various lighting condition. The image size is downsampled from 198×168 to 48×42 .

The CMU Multi-PIE database contains face images captured in 4 sessions with variations in illumination, expression and pose. We use the frontal images with neutral expression of the first 105 subjects of Session 1 in the experiments. Images are cropped based on the eye locations provided by [145] and down-sampled to 50×40 pixels.

Figure 5.5: Face samples from the Extended Yale B.

Figure 5.6: Face samples from the AR database.

For AR database, we use a subset including 50 male subjects and 50 female subjects. For each subject, 26 face images are taken in two separate sessions, that is, each session has 13 samples. These two sessions are with the same expression, illumination and disguise variation. However, Session 2 is captured one week later than Session 1. Fig. 5.6 shows images of a typical session from AR database. All images are resized from 165×120 to 55×40 .

The CMU PIE database has 41,368 facial images of 68 subjects. The face images were acquired by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. In the experiment, we choose the images captured from the frontal poses (C27) and use all the images under different illuminations and expressions. As a result, for each of 68 subjects, there are 49 facial images. Some facial images with illumination and expression changes are shown in Fig. 5.7. All images are with the size of 45×45 .

Figure 5.7: Face samples from the CMU PIE database.

5.4.1 Face Recognition with Uncorrupted Training Samples

This subsection tests the effectiveness of the proposed approaches with uncorrupted training data.

Experiment 1: for Extended Yale B database, we randomly select t from 64 illuminations as training data and the others are used for testing. We repeat this procedure 10 times for each t (6, 5, 4, 3). Table 5.1 details the averaged results of various methods with different training samples t . Much to our surprise, GL, which attempts to solve the problem of SRC, underperforms SRC in all scenarios. Among the holistic approaches, the proposed CSR achieves the best recognition rates, which is however worse than the patch based methods due to the extreme illuminations and insufficient training data. Although GL underperforms SRC all the time, the proposed GL-CP performs better than SRC-P in the cases of 4 and 3 training samples thanks to the collaborative patch scheme. WGSR consistently achieves the second best accuracies with the gains over SRC-P between 1.79% and 5.96%. The proposed CSR-CP performs the best and its improvement over SRC-P ranges from 4.85% to 14.8%.

Experiment 2: to further verify the performances of various methods in tackling the extreme illuminations, another experiment is done on Multi-PIE database. Following the procedure in [164], 18 flash-only images are generated as the differences between flash images (illuminations $\{1-18\}$) and non-flash image (illumination $\{0\}$) of Multi-PIE database as shown

Table 5.1: Recognition Rate on Uncorrupted Extended Yale B.

number t	6	5	4	3
LRC [71]	81.63%	78.14%	68.14%	60.69%
SRC [67]	82.97%	79.56%	74.0%	66.45%
CRC [165]	83.54%	81.41%	73.67%	68.47%
GL [96]	82.37%	78.84%	72.97%	66.2%
Proposed CSR	85.5%	82.12%	75.84%	69.6%
LRC-P [71]	79.06%	75.89%	68.21%	59.5%
SRC-P [67]	90.55%	89.38%	84.25%	74.6%
RCR [144]	77.67%	73.89%	69.17%	60.39%
WGSR [157]	<u>92.74%</u>	<u>91.17%</u>	<u>88.04%</u>	<u>80.56%</u>
Proposed GL-CP	89.44%	88.23%	87.62%	78.05%
Proposed CSR-CP	<u>95.4%</u>	<u>94.91%</u>	<u>93.71%</u>	<u>89.08%</u>

Figure 5.8: Face samples with illumination variation from CMU Multi-PIE database.

Figure 5.9: Recognition rate on Multi PIE Database.

in Fig. 5.8. For each of the 105 subjects, t images are randomly selected for training and the other $18 - t$ images are used as testing data. The averaged recognition rates over 10 runs are plotted against the reduced t in Fig. 5.9. It can be seen that the proposed CSR consistently outperforms the second best holistic approaches RCR. Although the recognition accuracies of GL are similar or even inferior to those of SRC, by using of collaborative patch scheme, the proposed GL-CP outperforms SRC-P and consistently perform the second best. When there are equal or more than 3 training samples per subject, the proposed CSR-CP achieves almost perfect results (i.e., greater than 99%). The maximum accuracy gain of the proposed CSR-CP over SRC-P reaches about 9% at 2 training samples per subject.

Experiment 3: for each subject of AR database, t (6, 5, 4, 3) of 7 undisguised samples from Session 1 are used for training and all 7 undisguised samples from Session 2 are used for testing. With each t , we randomly select 10 times. The averaged results are recorded in Table 5.2. Among all holistic approaches, the proposed CSR achieves the best recognition rates where

Table 5.2: Recognition Rate on AR Face Database.

number t	6	5	4	3
LRC [71]	72.82%	67.38%	60.52%	53.65%
SRC [67]	91.7%	89.84%	87.5%	81.55%
CRC [165]	94.13%	92.56%	91.13%	85.55%
GL [96]	92.42%	90.27%	88.70%	83.4%
Proposed CSR	94.28%	<u>93.71%</u>	<u>93.13%</u>	<u>89.27%</u>
LRC-P [71]	69.24%	67.24%	63.81%	56.51%
SRC-P [67]	92.7%	89.13%	87.72%	79.11%
RCR [144]	92.18%	90.94%	87.36%	83.88%
WGSR [157]	92.7%	90.27%	88.13%	82.40%
Proposed GL-CP	<u>94.56%</u>	93.42%	90.7%	85.98%
Proposed CSR-CP	<u>96.71%</u>	<u>95.57%</u>	<u>94.28%</u>	<u>90.84%</u>

the performance gain increases with decreasing number of training data. For LRC and SRC, their patch based versions cannot consistently outperform the corresponding holistic ones. In contrast, the proposed CSR-CP and GL-CP perform better than the proposed CSR and GL for all different t , respectively. Again, the proposed CSR-CP consistently achieves the best results.

Experiment 4: for each subject of CMU PIE dataset, t (4, 3, 2) facial images are randomly selected for training and all the other samples are used for testing. The averaged recognition rates over 10 runs are shown in Table 5.3. As the variation of this dataset is similar to that of Experiment 3, which includes only small illumination and facial expression, comparable results are obtained. Although SRC-P performs inferior to SRC due to the limited discriminative information of each patch, the proposed CSR-CP and GL-CP consistently outperform their corresponding holistic versions. For all number of training data, the proposed CSR-CP and CSR perform as the top 2 approaches.

5.4.2 Face Recognition with Occluded Training Samples

This subsection tests the effectiveness of various approaches on training data with disguise.

Table 5.3: Recognition Rate on CMU PIE Database.

number t	4	3	2
LRC [71]	76.77%	66.56%	49.64%
SRC [67]	92.12%	87.87%	82.43%
CRC [165]	92.68%	88.27%	83.62%
GL [96]	92.05%	87.78%	82.74%
Proposed CSR	<u>93.29%</u>	<u>89.79%</u>	<u>85.31%</u>
LRC-P [71]	75.17%	69.25%	64.92%
SRC-P [67]	90.54%	86.46%	82.37%
RCR [144]	90.69%	84.70%	81.30%
WGSR [157]	92.31%	87.26%	80.96%
Proposed GL-CP	92.38%	88.19%	84.47%
Proposed CSR-CP	<u>94.24%</u>	<u>90.34%</u>	<u>88.01%</u>

Table 5.4: Recognition rate on AR database with only disguised training samples.

LRC	SRC	CRC	GL	CSR	LRC-P	SRC-P	RCR	WGSR	GL-CP	CSR-CP
59.2%	83.1%	91.36%	88.49%	92.12%	75.72%	88.74%	91%	86.23%	<u>93.74%</u>	<u>96.75%</u>

Experiment 5: for each subject of AR database, all 6 disguised face images (i.e., the second row of Fig. 5.6) from Session 1 are used for training and 8 face images with illumination variations (i.e., the 1st, 5th, 6th, and 7th samples of the first row of Fig. 5.6) from both Session 1 and Session 2 are used for testing. Table 5.4 plots the performances of various methods in this scenario. The proposed CSR significantly outperforms SRC with the accuracy gain of 9.02% while that of GL is only 5.39%. The proposed CSR-CP achieves the accuracy gain of 8.01% from SRC-P.

The running time of each approach is evaluated under the Matlab programming environment and on a desktop of 3.5GHZ with 16G RAM. It is 0.02s for LRC and LRC-P, 0.12s for GL, 0.34s for SRC, 0.32s for SRC-P, 0.46s for WGSR, 0.88s for GL-CP, 0.03 for CRC, 0.32 for RCR, 6.65s for CSR and 19.47s for CSR-CP.

To analyze the convergence of the proposed CSR method, we plot the curve of first con-

Figure 5.10: The convergence curve of $\|y - Ax - e\|_\infty$ versus iteration number.

vergence condition of Step 7 Algorithm 3 in Fig. 5.10. It can be seen that $\|y - Ax - e\|_\infty$ decreases as the iteration number increases and converges after 650 iterations.

Experiment 6: for each subject of AR database, t (6, 5, 4, 3) facial images are randomly selected from all 13 images of Session 1 for training, and all 13 samples from Session 2 are used for testing. With each t , we repeat this procedure 10 times and record the averaged results in Table 5.5. It should be noted that the testing data of every subject have two different disguise types, both of which could be absent in its training data but could present in some other subjects. Table 5.5 shows that GL and CRC visibly outperform SRC. Much more significant gains over SRC are achieved by the proposed CSR. The performance of SRC-P is slightly better than that of SRC in this challenge scenario. However, by employing the proposed collaborative patch scheme, the accuracies of both proposed GL-CP and CSR-CP are visibly better than their corresponding holistic ones in all number of training data. The proposed CSR-CP again outperforms the others consistently. The accuracy gain over SRC-P reaches 18.34% at $t = 3$.

Table 5.5: Recognition Rate on Disguised AR Face Database.

number t	6	5	4	3
LRC [71]	48.77%	44.76%	36.67%	27.81%
SRC [67]	78.43%	74.5%	68.95%	59.86%
CRC [165]	83.98%	80.20%	74.37%	68.03%
GL [96]	81.66%	78.97%	72.65%	65.33%
Proposed CSR	<u>87.44%</u>	<u>83.74%</u>	<u>79.97%</u>	<u>72.57%</u>
LRC-P [71]	58.86%	58.17%	47.82%	46.46%
SRC-P [67]	79.66%	74.8%	69.88%	60.55%
RCR [144]	83.44%	80.66%	75.73%	71.11%
WGSR [157]	81.74%	78.58%	73.65%	64.87%
Proposed GL-CP	82.51%	79.28%	75.35%	69.8%
Proposed CSR-CP	<u>90.37%</u>	<u>88.29%</u>	<u>84.21%</u>	<u>78.89%</u>

Figure 5.11: Face samples partitioned into different numbers of patch. (a) 2×2 ; (b) 3×2 ; (c) 4×2 ; (d) 5×2 (e) 6×2

5.4.3 Face Recognition with Different Patch Numbers

This subsection tests the effectiveness of various approaches with different number of patch L . The images are partitioned into L patches, here L is set to 4 (2×2), 6 (3×2), 8 (4×2), 10 (5×2), or 12 (6×2) as shown in Fig. 5.11.

Experiment 7: the data setting of Experiments 6 with 4 training samples is used to test various patch based approaches with different number of patch. Fig. 5.12 plots the performances of averaged results of 10 runs. It shows that the results of SRC-P and LRC-P vary visibly as the number of patch increases. SRC-P delivers recognition rates ranging from 63.25% to 72.42%, while the performance of LRC-P is from 45.53% to 51.93%. The approaches em-

Figure 5.12: Recognition rate on AR database with different numbers of patch.

playing relations between patches are less sensitive to the number of patch. The difference of maximum and minimum accuracies of RCR, WGSR, the proposed GL-CP and CSR-CP are 1.36%, 2.24%, 3.16% and 1.78%, respectively. Fig. 5.12 also demonstrates that the proposed CSR-CP significantly outperforms all other algorithms consistently for all numbers of patch.

5.5 Summary

SRC considers the face recognition problem as a sparse linear representation of the query image. Ideally, all significant representation coefficients should be associated to the correct subject. However, as face images of different subjects could be similar and correlated, the sample-wise sparse representation that ignores the identity information of training samples tends to employ training samples of many subjects. This results in misclassification. GL attempts to solve this problem by minimizing the so-called $\ell_{2,1}$ -norm of the representation coefficients. Although the minimization of the ℓ_1 -norm component leads to a group sparsity, the minimization of the ℓ_2 -

norm component may result in the attained representation deviating from the desired solution as the optimal representation by training samples of the correct subject may not necessarily be dense. The proposed CSR alleviates the problems of SRC and GL. It seeks an optimum representation of the query image by minimizing the number of selected classes of training data.

Patch based approaches are effective to handle the corrupted images and extreme image variations at a price of reducing the discriminative information in the training data. Different from the conventional patch based methods that optimize each patch separately, the proposed CSR-CP approach optimizes all patches together to seek a collaborative patch group-wise sparse representation. It inherits the merits of the patch based approaches in dealing with the corrupted or extremely variate images and alleviates their problems in reducing the discriminative information. As a result, a more reliable and discriminative representation is achieved by the proposed CSR-CP method. Extensive experiments on several benchmark databases verify the merits and significance of the proposed CSR-CP approach.

Chapter 6

Sparse and Dense Hybrid Representation via Dictionary Decomposition for Face Recognition

Sparse representation provides an effective tool for classification under the conditions that every class has sufficient representative training samples and the training data are uncorrupted. These conditions may not hold true in many practical applications. Face identification is an example where we have a large number of identities but sufficient representative and uncorrupted training images cannot be guaranteed for every identity. A violation of the two conditions leads to a poor performance of the sparse representation-based classification (SRC). This chapter addresses this critical issue by analyzing the merits and limitations of SRC. A sparse- and dense-hybrid representation (SDR) framework is proposed in this chapter to alleviate the problems of SRC. We further propose a procedure of supervised low-rank (SLR) dictionary decomposition to facilitate the proposed SDR framework. In addition, the problem of the corrupted training data is also alleviated by the proposed SLR dictionary decomposition. The application of the proposed SDR-SLR approach in face recognition verifies its effectiveness and advancement to the field. Extensive experiments on benchmark face databases demonstrate that it consistently outperforms the state-of-the-art sparse representation based approaches and the performance

gains are significant in most cases.

6.1 Introduction

Different from the traditional classifiers (e.g., nearest neighborhood, the nearest feature line, the nearest plane and the nearest subspace), SRC applies sparse representation of the query image by training samples of all classes to design the classifier for face recognition. Some promising results have been delivered thanks to its properties, which attract great interest in further research on SRC. Although many extensions of SRC have been proposed, they have not solved the two fundamental constraints of SRC. One is the carefully controlled training images and the other is the sufficient training samples of each class. A violation of these two conditions results in poor performance of the sparse representation-based classification [148, 149]. The first constraint makes SRC based approaches do not perform well for the corrupted training data caused by outlier pixels and occlusions. The second constraint limits their applications to large scale identification problems where the training data contains large number of identities but sufficient representative images for every identity cannot be guaranteed. Therefore, it is not a surprise that, despite of the impressive results of SRC and many of its variants and extensions developed, a number of works [97, 100, 166, 167] show questions and doubts about the effectiveness of SRC for image classification.

The both fundamental constraints of the sparse representation-based classification are related to the training data. As SRC directly applies training images as the dictionary for the sparse representation, questions may arise if dictionary learning [168] can help alleviate these two problems of SRC. The objective of the general dictionary learning [106, 111] is to find different optimal dictionaries in representing different target data. The atoms in a dictionary computed from the training database in general capture the most important constitutive component of the

target data. However, there is no good reason why the set of atoms is also best for differentiating different classes in the target data. The discriminative dictionary learning [108, 110, 112, 169] makes a compromise between the data representation and the class separation in finding the optimal atoms of dictionary. This, however, also does not solve the problem of SRC when not all classes have sufficient representative training samples. In addition, all these dictionary learning approaches do not work well if the training images are heavily corrupted by the outlier pixels and occlusions [168]. Instead of the conventional dictionary learning approaches, it seems that we need separate the image corruptions of the training samples from the dictionary.

To overcome the first constraints of SRC – corrupted training data, low-rank matrix recovery, also called robust PCA (RPCA) [117], provides a tool to separate outlier pixels and occlusions from the training images. A sub-dictionary is learnt from each class separately by decomposing the training data of each class into a low-rank matrix and a sparse matrix in [114, 115]. However, optimizing sub-dictionaries to be low-rank for each class might reduce the diversity of atoms within the sub-dictionary and hence might decrease the dictionary’s representation power [116]. Following the idea of subspace clustering by low-rank representation [118], low-rank representation (coefficient matrix of the dictionary) is learnt for image classification [116]. However, the natural cluster structure of the training data may not coincide with the class (identity) structure for a face identification problem due to the small differences between the face identities and the large variations of face images of the same identity. As a result, in seeking a low rank dictionary or a low rank representation, these methods may undesirably remove some components of face identity and/or some face variations needed to represent the query images.

There are some attempts to alleviate the second problem of SRC – lack of variations in the training images for some classes. The extended SRC (ESRC) [94] creates an intra-class variation matrix by subtracting a natural or prototype or centroid image of each class from training images of the same class. The variation matrix is then appended to the raw training data

matrix as the second part of the dictionary. The classification only utilizes the representation coefficients of the first part. Although the query images can be better represented by some variations of other classes in the second part of the dictionary, the same variations are also contained in the first part. Therefore, the introduced second part of the dictionary may not well solve the problem of the first part. This leads the same authors of ESRC to further propose a superposed SRC (SSRC) [95] by replacing the first part with a natural or prototype or centroid image of each class. However, this brings another problem that the identity of the query image is represented and determined by only a single image, which may result in unstable or unreliable classification.

This work aims to solve the two fundamental limitations of SRC – the lack of variations in the training images for some classes and the corrupted training data. We propose a hybrid representation of the query image by a sparse combination of a class-specific dictionary and a dense combination of a common intra-class variation dictionary. Thus, a query image is better represented with the collaboration of image variations of other classes, which leads the sparse representation part better to coincide with the specified class membership. Towards this end, we further propose a procedure to decompose the raw training data into a class-specific dictionary, a common intra-class variation dictionary and a sparse corruption matrix. In this way, the corruption matrix will also be better separated from two lower-rank matrices than from the higher-rank mixture of them. In implementation, the proposed approach needs solve a nuclear norm regularized optimization, which is a convex minimization problem solvable in polynomial time.

6.2 Sparse and Dense Hybrid Representation

An image of m pixels is arranged in a column vector. Let $\mathbf{A}_i \in \mathbb{R}^{m \times n_i}$ stack n_i m -dimensional training samples of the i th class. An unlabeled query image $\mathbf{y} \in \mathbb{R}^m$ is represented or approximated by a linear combination of the training samples from a class as

$$\mathbf{y} = \mathbf{A}_i \mathbf{x}_i + \mathbf{e}_i, \quad i = 1, 2, \dots, c, \quad (6.1)$$

where $\mathbf{x}_i$ is a coefficient vector associated with the training samples of class i and $\mathbf{e}_i$ is the approximation error. It is widely assumed that images of a specific class lie in a linear subspace [6, 70]. Thus, the linear regression classifier (LRC) [71, 170] assigns the class label of $\mathbf{y}$ to class i that produces the smallest error $\|\mathbf{e}_i\|_2$. This is in fact equivalent to a nearest subspace [69, 143, 171] classifier as $\|\mathbf{e}_i\|_2$ is the distance from $\mathbf{y}$ to the subspace spanned by columns of $\mathbf{A}_i$.

There is a problem if $n_i \ll m$ or some classes do not have representative training samples. The variation between the query image and the training samples of the same class could be larger than those of some other classes, which results in misclassification. Even for a large number of training samples of each class, there is another problem that errors of some classes could be so small that the classification based on them is unreliable. This can be understood by the fact that there exists a representation $\mathbf{x}_i$ leading to $\mathbf{e}_i = 0$ for any class as long as its training samples are not linear dependent and $n_i \geq m$.

Now consider a collaborative representation of the query image by training samples of all classes as

$$\mathbf{y} = \mathbf{A} \mathbf{x} + \mathbf{e}, \quad (6.2)$$

where $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_c] \in \mathbb{R}^{m \times n}$ stacks training samples of all c classes, $n = \sum_{i=1}^c n_i$. If

$\text{rank}(\mathbf{A}) = m = n$, we can have an unique perfect representation $\mathbf{x}$ that leads to $\mathbf{e} = 0$ for any query image $\mathbf{y}$. If the dictionary $\mathbf{A}$ is over-complete, i.e. $n > m$, the perfect representation $\mathbf{x}$ is not unique, i.e. there are an infinite number of solutions $\mathbf{x}$ that lead to $\mathbf{e} = 0$. Thus, a perfect representation with the sparsity constraint on $\mathbf{x}$ can be found. We can achieve a more sparse solution $\mathbf{x}$, or get it even on an under-complete dictionary, $n < m$, by relaxing the requirement for the perfect representation. Therefore, we look for a sparse linear representation $\mathbf{x}$ of $\mathbf{y}$ over $\mathbf{A}$ with a bounded energy of the representation error $\|\mathbf{e}\|_2^2 < \varepsilon$.

The sparsity of $\mathbf{x}$ is measured by the number of its nonzero elements, i.e. ℓ_0 -norm of $\mathbf{x}$, $\|\mathbf{x}\|_0$. This optimization problem is formulated as

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \text{ s.t. } \|\mathbf{e}\|_2^2 < \varepsilon, \text{ where } \mathbf{e} = \mathbf{y} - \mathbf{Ax}. \quad (6.3)$$

Although this minimization problem is NP-hard, the compressed sensing theory reveals that, if its solution is sufficient sparse, it is equivalent to the following convex relaxation ℓ_1 -norm minimization [84]

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \beta \|\mathbf{e}\|_2^2, \text{ where } \mathbf{e} = \mathbf{y} - \mathbf{Ax}, \quad (6.4)$$

where β is a constant for a compromise between the sparsity of $\mathbf{x}$ and the representation error $\|\mathbf{e}\|_2^2$.

To enable a robust representation of heavily corrupted query image by outlier pixels and occlusions, SRC [67, 148] replaces the ℓ_2 -norm of the representation error by ℓ_1 -norm in the optimization

$$\min_{\mathbf{x}, \mathbf{e}} \|\mathbf{x}\|_1 + \beta \|\mathbf{e}\|_1, \text{ s.t. } \mathbf{y} = \mathbf{Ax} + \mathbf{e}. \quad (6.5)$$

From (6.2) and (6.5) we see three merits of SRC: the collaborative representation using training samples of all classes $\mathbf{A}$; the sparse representation code $\mathbf{x}$; and the ℓ_1 -norm minimization

of the representation error $\|e\|_1$. These merits make the sparse representation based approaches a great success in image restoration [58, 60] and face recognition [67, 148, 149] in some scenarios. However, SRC does not perform well for corrupted training data or if the sufficient representative training samples are not provided for every class. Thus, a number of works [97, 100, 166, 167] show questions and doubts about the effectiveness of sparse representation for image classification.

We argue that the sparse representation directly coincides with the general objective of the classification because the desired output or target of a classification system is 1 for correct class and 0 for all others, which is exactly a sparse code of the query image. In addition, the collaborative representation compensates the limited number of representative samples of a single class, though only partially due to the sparse constraint. The optimization of the sparse coefficients lets every training sample compete against the others to win its representation share of the query image. This is a good discriminating process.

Problem is that images contain not only the identity information but also much other information such as age, gender, ethnic, expression and illumination. Why must the significant coefficients in $\mathbf{x}$ be won by the samples of the same identity as the query image, not of the same expression or the same illumination? Neither (6.2) nor (6.5) receives any information about the class label assignment of the training data that specifies the particular classification task. If we can separate the class-specific information from others, the sparse representation of the former will fully coincides with the particular classification target specified by the class labels of the training data.

Therefore, we propose to decompose an image $\mathbf{y}$ into three components as

$$\mathbf{y} = \mathbf{d} + \mathbf{b} + \mathbf{s}, \quad (6.6)$$

where $\mathbf{d}$ is the class-specific component, $\mathbf{b}$ the non-class-specific variations and $\mathbf{s}$ contains random sparse noise or image corruption. If a dictionary $\mathbf{D}$ that only contains class-specific component of the image is available, we can apply SRC as

$$\mathbf{d} = \mathbf{D}\mathbf{x} + \mathbf{e}_d. \quad (6.7)$$

The sparse vector $\mathbf{x}$ of (6.7) will directly coincide with the class label vector of the query image $\mathbf{y}$ while that of (6.2) only coincides with the natural image clusters as $\mathbf{d}$ and $\mathbf{D}$ only have the class-specific components but $\mathbf{y}$ and $\mathbf{A}$ contain all other image information.

Although it might be possible to separate the class-specific component $\mathbf{d}$ from $\mathbf{y}$ based on the human knowledge in some applications, it is very difficult if not impossible for many computer vision tasks such as face recognition. Given a labeled training database $\mathbf{A}$, it is possible to decompose it into a class-specific dictionary $\mathbf{D}$, a non-class-specific dictionary $\mathbf{B}$ and a random sparse noise $\mathbf{E}$ based on machine learning. This will be presented in the next section.

However, it is very difficult if not impossible to separate the class-specific component $\mathbf{d}$ from a single unknown query image $\mathbf{y}$. To circumvent this problem, we also represent the non-class-specific component $\mathbf{b}$ by the dictionary $\mathbf{B}$ as

$$\mathbf{b} = \mathbf{B}\boldsymbol{\gamma} + \mathbf{e}_b. \quad (6.8)$$

The only purpose of the representation (6.8) is to make the representation error $\mathbf{e}_b$ as small as possible. Therefore, it is not necessary to put the sparse constraint on $\boldsymbol{\gamma}$ and hence (6.8) should be in general a dense representation.

The summation of (6.7), (6.8) and $\mathbf{s}$ yields the proposed sparse- and dense-hybrid representation (SDR) of the query image

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{B}\boldsymbol{\gamma} + \mathbf{e}, \quad (6.9)$$

where $\mathbf{e} = \mathbf{e}_d + \mathbf{e}_b + \mathbf{s}$ is the combined representation error. Fig. 6.1 illustrates a face image projected to the identity subspace (spanned by $\mathbf{d}$), the common variation subspace (spanned by $\mathbf{B}$) and its sparse residual.

Figure 6.1: Image decomposition: a face image projected to the class-specific subspace, the non-class-specific subspace and their residual random sparse noise.

To be robust to the heavily corrupted query image by outlier pixels and occlusions, ℓ_1 -norm minimization of the representation error $\mathbf{e}$ is applied. The solution of the proposed SDR, $\mathbf{x}$, $\boldsymbol{\gamma}$ and $\mathbf{e}$ is obtained by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{x}, \boldsymbol{\gamma}, \mathbf{e}} \quad & \|\mathbf{x}\|_1 + \sigma \|\boldsymbol{\gamma}\|_2^2 + \beta \|\mathbf{e}\|_1 \\ \text{s.t.} \quad & \mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{B}\boldsymbol{\gamma} + \mathbf{e}. \end{aligned} \quad (6.10)$$

The optimization problem (6.10) can be solved by the Augmented Lagrange Multiplier

(ALM) scheme [146]. The ALM function for (6.10) is derived as:

$$\|\mathbf{x}\|_1 + \sigma\|\boldsymbol{\gamma}\|_2^2 + \beta\|\mathbf{e}\|_1 + \frac{\xi}{2}\|\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{B}\boldsymbol{\gamma} - \mathbf{e}\|_2^2 + \boldsymbol{\phi}^T(\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{B}\boldsymbol{\gamma} - \mathbf{e}), \quad (6.11)$$

where $\boldsymbol{\phi}$ is a vector of Lagrange multipliers and ξ a penalty parameter. Algorithm 4 summarizes the solution to problem (6.10). The subproblem for $\mathbf{e}$ in step 1 can be solved by the soft-thresholding operator [79], $S(\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{B}\boldsymbol{\gamma} + \frac{1}{\xi}\boldsymbol{\phi}, \frac{\beta}{\xi})$. Here $S(\cdot, \cdot)$ is defined element-wise as follows:

$$S(\mathbf{t}, \omega)_q = \text{sign}(t_q) \cdot \max\{|t_q| - \omega, 0\}, \text{ for } q = 1, \dots, m. \quad (6.12)$$

The result of $\mathbf{x}$ can be achieved by ℓ_1 -norm minimization and $\boldsymbol{\gamma}$ has a closed-form solution.

Figure 6.2: Comparison of SRC (1th row) and the proposed SDR (2nd row) on AR database A of 100 persons with 3 images per person: (a) query image; (b) SRC coefficients of $\mathbf{A}$ and SDR coefficients of $\mathbf{D}$ and $\mathbf{B}$; (c) top 16 images in $\mathbf{A}$ of SRC and in $\mathbf{D}$ and $\mathbf{B}$ of SDR; (d) residuals. Sparse coefficients of the correct person are in red lines and its training samples are in yellow rectangles. Images are normalized into the same mean and ℓ_2 -norm.

Let $\mathbf{L}_i \in \mathbb{R}^{n \times n}$ be a class-label matrix of the training data $\mathbf{A}$ for class i , its element $\mathbf{L}_i(k, k) = 1$ if the k th training sample (the k th column of $\mathbf{A}$) originates from class i and

Algorithm 4: Solving Problem(6.10) by Inexact ALM

Input: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{y}$, parameter β and σ .Initialize: $\mathbf{x} = 0$, $\boldsymbol{\gamma} = 0$, $\mathbf{e} = 0$, $\boldsymbol{\phi} = 0$, $\xi = 1$, $\xi_{max} = 10^6$, $\rho = 1.5$, and $\epsilon = 10^{-6}$.**while** not converged **do**1. fix the others and update $\mathbf{e}$ by

$$\mathbf{e} = \arg \min_{\mathbf{e}} \frac{\beta}{\xi} \|\mathbf{e}\|_1 + \frac{1}{2} \|\mathbf{e} - (\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{B}\boldsymbol{\gamma} + \frac{1}{\xi}\boldsymbol{\phi})\|_2^2$$

2. fix the others and update $\boldsymbol{\alpha}$ by

$$\mathbf{x} = \arg \min_{\mathbf{x}} \|\mathbf{x}\|_1 + \frac{\xi}{2} \|\mathbf{D}\mathbf{x} - (\mathbf{y} - \mathbf{B}\boldsymbol{\gamma} - \mathbf{e} + \frac{1}{\xi}\boldsymbol{\phi})\|_2^2$$

3. fix the others and update $\mathbf{x}$ by

$$\boldsymbol{\gamma} = \xi(2\sigma\mathbf{I} + \xi\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{e} + \frac{1}{\xi}\boldsymbol{\phi})$$

4. update the multipliers

$$\boldsymbol{\phi} = \boldsymbol{\phi} + \xi(\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{B}\boldsymbol{\gamma} - \mathbf{e})$$

5. update ξ by $\xi = \min(\rho\xi, \xi_{max})$

6. check the convergence conditions:

$$\|\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{B}\boldsymbol{\gamma} - \mathbf{e}\|_2^2 < \epsilon$$

end**Output:** $\mathbf{x}$, $\boldsymbol{\gamma}$ and $\mathbf{e}$

all other elements of $\mathbf{L}_i$ are zero. The representation of the query image $\mathbf{y}$ by the class-specific component of class i and the non-class-specific component of all classes collaboratively is then

$$\mathbf{y} = \mathbf{D}\mathbf{L}_i\mathbf{x} + \mathbf{B}\boldsymbol{\gamma} + \mathbf{e}_i. \quad (6.13)$$

As discussed earlier, the sparse vector $\mathbf{x}$ coincides with the class-label vector (the sum of $\mathbf{L}_i$ over all columns). Therefore, the representation error $\mathbf{e}_i$ will be the closest to $\mathbf{e}$ if the query image $\mathbf{y}$ originates from class i . The class-wise representation residual is defined by

$$r_i(\mathbf{y}) = \|\mathbf{e} - \mathbf{e}_i\|_2 = \|\mathbf{D}(\mathbf{I} - \mathbf{L}_i)\mathbf{x}\|_2 = \|\mathbf{D}\mathbf{x} - \mathbf{D}\delta_i(\mathbf{x})\|_2, \quad (6.14)$$

where $\mathbf{I}$ is an identity matrix and $\delta_i(\mathbf{x})$ only retains the coefficients associated with the i th class. The query image $\mathbf{y}$ is classified into the class that produces the minimum residual $r_i(\mathbf{y})$.

Sparse representation (6.2) coincides with the general objective of classification: degenerate all query images to only two different states: one for the correct class and the other for all other classes. However, the competition, in which a training sample gains its share to represent the query image by (6.2) and (6.5), is unsupervised. As a result, all information contained in the query image and training samples participates in this competition. Thus, there is no reason why the nonzero significant coefficients should coincide with a specific one of many possible class arrangements that (6.2) and (6.5) do not know. For many computer vision tasks, such as face recognition, the class-specified information is only a very small portion of the whole information carried by an image. Therefore, SRC can only work well if all non-class-specific information is nulled out by sufficient representative training samples for every class that enclose all non-class-specific information. This problem is visible in the first row of Fig. 6.2, which shows an example of SRC where the top 3 significant coefficients are not for the training samples of the same identity as the query image.

The proposed sparse- and dense-hybrid representation (SDR) alleviates this problem. To represent a query image, every training sample only uses its class-specific component to compete against the others (through a sparse minimization) collaboratively with the non-class-specific component of all training samples (through a dense representation). This makes the sparse code of the proposed SDR completely coincide with the specifically assigned class membership. It overcomes a fundamental limitation of SRC that sufficient representative training samples are required for every class. If a class does not have sufficient training samples to represent some variations (non-class-specific component) of its query image, they are represented by the non-class-specific component of other classes and their class memberships do not participate in the classification competition. This is visible in the second row of Fig. 6.2, which shows an example of the proposed SDR where the top and the fourth significant coefficients are for the training samples of the same identity as the query image.

6.3 Dictionary Decomposition

The sparse- and dense-hybrid representation (SDR) presented in the last section requires a class-specific dictionary $\mathbf{D}$ and a non-class-specific dictionary $\mathbf{B}$ decomposed from training data $\mathbf{A}$. Towards this end, we apply the proposed conceptual SDR model (9) to every training image in $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_k, \dots, \mathbf{a}_n]$ as

$$\mathbf{a}_k = \mathbf{D}\mathbf{x}_k + \mathbf{B}\boldsymbol{\gamma}_k + \mathbf{e}_k, \quad k = 1, \dots, n. \quad (6.15)$$

Here we set $\mathbf{x}_k$ be the sparsest vector as $\mathbf{x}_k(j) = 1$ if $j = k$ and $\mathbf{x}_k(j) = 0$ if $j \neq k$ for $j, k = 1, 2, \dots, n$. This is plausible as full information of training sample $\mathbf{a}_k$ can be utilized in the right side of equation (6.15). Stacking vectors $\mathbf{a}_k, \mathbf{x}_k, \boldsymbol{\gamma}_k, \mathbf{e}_k$ of all different k in (6.15) into

respective matrices yields

$$\mathbf{A} = \mathbf{D} + \mathbf{B}\mathbf{\Gamma} + \mathbf{E} \quad (6.16)$$

The sparse coefficients of $\mathbf{D}$ disappear because $[\mathbf{x}_1, \dots, \mathbf{x}_n]$ forms an identity matrix.

To realize the proposed SDR framework, we further propose to decompose the training data $\mathbf{A}$ according to the model (6.16) into a class-specific dictionary $\mathbf{D}$, a non-class-specific dictionary $\mathbf{B}$ and a sparse noise or corruption $\mathbf{E}$, where $\mathbf{\Gamma}$ helps the dictionary $\mathbf{B}$ have good representation (prediction) power. If the training data are meaningful, we can reasonably assume $\mathbf{A}$ be a full rank matrix, i.e. $\text{rank}(\mathbf{A}) = \min(m, n)$. This is true with the existence of the random noise in the training data. As the training data $\mathbf{A}$ does not contain all kinds but only a particular type of images – human faces for example, $\mathbf{A} - \mathbf{E}$ should be a low rank matrix. Consequently, $\mathbf{D}$, $\mathbf{B}\mathbf{\Gamma}$ and $\mathbf{B}$ are all low rank matrices. Therefore, the dictionary decomposition (6.16) is regularized by

$$\begin{aligned} \min_{\mathbf{D}, \mathbf{B}, \mathbf{\Gamma}, \mathbf{E}} \quad & \text{rank}(\mathbf{D}) + \lambda \text{rank}(\mathbf{B}) + \tau \|\mathbf{\Gamma}\|_F^2 + \eta \|\mathbf{E}\|_0 \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{D} + \mathbf{B}\mathbf{\Gamma} + \mathbf{E}. \end{aligned} \quad (6.17)$$

However, the minimization of the ranks of $\mathbf{D}$ and $\mathbf{B}$ and the number of nonzero elements in $\mathbf{E}$ is a highly non-convex optimization problem, which is difficult to solve. Fortunately, it is proven in [117] that the solution of low-rank matrix recovery problem can be well approximated by replacing the rank operator with the nuclear norm $\|\cdot\|_*$ and replacing the ℓ_0 -norm with the ℓ_1 -norm, which turn it into a convex optimization problem. It is proven that this convex relaxed optimization well recovers the low-rank matrix and the sparse error if the rank of the matrix to be minimized is not too high and the respective error matrix is sparse [117]. A number of algorithms [117, 147, 172] are proposed to solve this convex relaxed problem. Therefore, we

relax the regularization of the dictionary decomposition to

$$\begin{aligned} \min_{\mathbf{D}, \mathbf{B}, \mathbf{\Gamma}, \mathbf{E}} \quad & \|\mathbf{D}\|_* + \lambda \|\mathbf{B}\|_* + \tau \|\mathbf{\Gamma}\|_F^2 + \eta \|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{D} + \mathbf{B}\mathbf{\Gamma} + \mathbf{E}. \end{aligned} \quad (6.18)$$

The squared Frobenius norm $\|\mathbf{\Gamma}\|_F^2$ is the sum of the squared ℓ_2 -norms, $\|\boldsymbol{\gamma}_k\|_2^2$. λ , τ and η are parameters to balance the minimization of the four terms.

The decomposition regularization (6.18) can be solved by minimizing two of the four unknowns with fixed others iteratively. Some initial values are required for the iterative minimization. The random sparse noise is initialized as a null matrix, $\mathbf{E} = \mathbf{0}$. Vector $\boldsymbol{\gamma}_k$ is a dense representation over the non-class-specific dictionary $\mathbf{B}$. We initialize it as the sparsest vector and let it approach a dense vector during the subsequent learning process. Thus, $\boldsymbol{\gamma}_k(j) = 1$ if $j = k$ and $\boldsymbol{\gamma}_k(j) = 0$ if $j \neq k$ for $j, k = 1, 2, \dots, n$. Thus, $\mathbf{\Gamma}$ is initialized as an identity matrix $\mathbf{\Gamma} = \mathbf{I}$. Although we decompose the training data $\mathbf{A}$ into four parts $\mathbf{D}$, $\mathbf{B}$, $\mathbf{\Gamma}$ and $\mathbf{E}$, only the two dictionaries $\mathbf{D}$, $\mathbf{B}$ are utilized by the proposed SDR. The above initialization of $\mathbf{E}$ and $\mathbf{\Gamma}$ yields $\mathbf{A} = \mathbf{D} + \mathbf{B}$. Thus, the dictionaries $\mathbf{D}$ and $\mathbf{B}$ have not lost any information from the training data $\mathbf{A}$ in the above initialization of $\mathbf{E}$ and $\mathbf{\Gamma}$.

As the decomposition regularization (6.18) involves two low-rank matrices $\mathbf{D}$ and $\mathbf{B}$ simultaneously, there is no reason why $\mathbf{D}$ must capture the class-specific information while $\mathbf{B}$ the non-class-specific one. Therefore, knowledge about the different roles of these four matrices in the proposed SDR must be applied to help decompose the training data $\mathbf{A}$ properly into meaningful $\mathbf{D}$, $\mathbf{B}$, $\mathbf{\Gamma}$ and $\mathbf{E}$ for the proposed SDR.

Towards this end, we try to allot all possible class-specific information we can have to $\mathbf{D} = [\mathbf{D}_1, \dots, \mathbf{D}_i, \dots, \mathbf{D}_c]$, $\mathbf{D}_i \in \mathbb{R}^{m \times n_i}$. If we let $\mathbf{D}_i$ be some kind of class-conditional center such as the mean or median of the training samples of class i , $\mathbf{D}$ should have captured

the most significant class-specific information. This is supported by the fact that all approaches of the linear discriminant analysis only utilize the class-conditional mean as the class-specific information. For a better representation power of $\mathbf{D}_i$, this work applies singular value decomposition (SVD) on every class-specific training data $\mathbf{A}_i$

$$\mathbf{A}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^T, \quad (6.19)$$

and lets $\mathbf{D}_i$ be the reconstructed images by the singular vectors corresponding to the largest singular value

$$\mathbf{D}_i = \mathbf{U}_i(1:m, 1) \mathbf{\Sigma}_i(1, 1) \mathbf{V}_i(1:n_i, 1)^T. \quad (6.20)$$

This makes $\mathbf{D}_i$ the lowest rank (rank 1) and the best approximation of $\mathbf{A}_i$ in terms of the Frobenius norm of the difference between $\mathbf{D}_i$ and $\mathbf{A}_i$. Therefore, $\mathbf{D}$ captures the most significant class-specific information with the lowest possible rank, $\text{rank}(\mathbf{D}) = c$. In other words, it allots the least amount of the most significant class-specific information of the training data $\mathbf{A}$ into $\mathbf{D}$. The matrix $\mathbf{B} = \mathbf{A} - \mathbf{D}$ then contains all non-class-specific information and the remaining class-specific information. Obviously, $\text{rank}(\mathbf{B}) = \min(m, n) - c$. In most applications we have in fact $\text{rank}(\mathbf{B}) \gg \text{rank}(\mathbf{D})$. Thus, the subsequent dictionary learning should target at transferring the remaining class-specific information from $\mathbf{B}$ to $\mathbf{D}$.

We first separate the remaining class-specific information from $\mathbf{B}$ by applying the decomposition regularization (6.18) with fixed $\mathbf{D}$ and $\mathbf{\Gamma}$

$$\begin{aligned} & \min_{\mathbf{B}, \mathbf{T}} \lambda \|\mathbf{B}\|_* + \delta \|\mathbf{T}\|_1 \\ & \text{s.t. } \mathbf{A} - \mathbf{D} = \mathbf{B}\mathbf{\Gamma} + \mathbf{T}. \end{aligned} \quad (6.21)$$

Symbols $\mathbf{E}$ and η of (6.18) are changed to $\mathbf{T}$ and δ in (6.21) to indicate that more component than the random sparse noise should be captured by $\mathbf{T}$ to enable the information transfer. This

can be achieved by assigning a smaller weight δ to $\|\mathbf{T}\|_1$.

The minimization of $\text{rank}(\mathbf{B})$ minimizes $\text{rank}(\mathbf{B}\mathbf{\Gamma})$ as $\text{rank}(\mathbf{B}\mathbf{\Gamma}) \leq \text{rank}(\mathbf{B})$. Thus, the optimization (6.21) decomposes $\mathbf{A} - \mathbf{D}$ into a sparse matrix $\mathbf{T}$ and a low rank matrix $\mathbf{B}\mathbf{\Gamma}$. It coincides with the dense representation part of the proposed SDR (6.9), which helps the non-class-specific dictionary $\mathbf{B}$ better represent (predict) unknown query image in SDR than directly taking $\mathbf{B}\mathbf{\Gamma}$ as the dictionary. Given a small parameter δ , the sparse matrix $\mathbf{T}$ will capture not only the random sparse noise but also the remaining class-specific information in $\mathbf{A} - \mathbf{D}$.

To make the problem (6.21) solvable, an auxiliary variable $\mathbf{J}$ is introduced to convert (6.21) to the following equivalent optimization problem:

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{J}, \mathbf{T}} \quad & \lambda \|\mathbf{J}\|_* + \delta \|\mathbf{T}\|_1 \\ \text{s.t.} \quad & \mathbf{A} - \mathbf{D} = \mathbf{B}\mathbf{\Gamma} + \mathbf{T}, \mathbf{B} = \mathbf{J}. \end{aligned} \quad (6.22)$$

This can be solved by the ALM method [146] that converts (6.22) to minimizing an unconstrained function:

$$\begin{aligned} \lambda \|\mathbf{J}\|_* &+ \delta \|\mathbf{T}\|_1 + \text{tr}(\mathbf{Y}_1^T (\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} - \mathbf{T})) \\ &+ \frac{\mu}{2} \|\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} - \mathbf{T}\|_F^2 + \text{tr}(\mathbf{Y}_2^T (\mathbf{B} - \mathbf{J})) \\ &+ \frac{\mu}{2} \|\mathbf{B} - \mathbf{J}\|_F^2, \end{aligned} \quad (6.23)$$

where $\mathbf{Y}_1, \mathbf{Y}_2$ are the Lagrange multipliers, and μ is a penalty parameter. Algorithm 5 summarizes the solution to problem (6.21). Step 1 is solved via the Singular Value Thresholding [173] as $\mathbf{J} = \mathbf{U}S(\mathbf{\Sigma}, \frac{\lambda}{\mu})\mathbf{V}^T$, where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD of $(\mathbf{B} + \mathbf{Y}_2/\mu)$ and the soft-thresholding operator $S(\cdot, \cdot)$ is extended to matrix $\mathbf{\Sigma}$ by applying it element-wise. The solution of Step 3 of Algorithm 5 is achieved by soft-thresholding as $S(\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} + \mathbf{Y}_1/\mu, \frac{\delta}{\mu})$.

Algorithm 5: Solving Problem(6.21) by Inexact ALM**Input:** $\mathbf{A}$, $\mathbf{D}$, $\mathbf{\Gamma}$ and parameter λ and δ .Initialize: $\mathbf{B} = \mathbf{J} = 0$, $\mathbf{T} = 0$, $\mathbf{Y}_1 = 0$, $\mathbf{Y}_2 = 0$, $\mu = 10^{-3}$, $\mu_{max} = 10^6$, $\rho = 1.5$, and $\epsilon = 10^{-6}$.**while not converged do**1. fix the others and update $\mathbf{J}$ by

$$\mathbf{J} = \arg \min_{\mathbf{J}} \frac{\lambda}{\mu} \|\mathbf{J}\|_* + \frac{1}{2} \|\mathbf{J} - (\mathbf{B} + \mathbf{Y}_2/\mu)\|_F^2$$

2. fix the others and update $\mathbf{B}$ by

$$\mathbf{B} = ((\mathbf{A} - \mathbf{D} - \mathbf{T})\mathbf{\Gamma}^T + \mathbf{J} + (\mathbf{Y}_1\mathbf{\Gamma}^T - \mathbf{Y}_2)/\mu)(\mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{I})^{-1}$$

3. fix the others and update $\mathbf{T}$ by

$$\mathbf{T} = \arg \min_{\mathbf{T}} \frac{\delta}{\mu} \|\mathbf{T}\|_1 + \frac{1}{2} \|\mathbf{T} - (\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} + \mathbf{Y}_1/\mu)\|_F^2$$

4. update the multipliers

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{Y}_1 + \mu(\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} - \mathbf{T}) \\ \mathbf{Y}_2 &= \mathbf{Y}_2 + \mu(\mathbf{B} - \mathbf{J}) \end{aligned}$$

5. update μ by $\mu = \min(\rho\mu, \mu_{max})$

6. check the convergence conditions:

$$\|\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} - \mathbf{T}\|_F^2 < \epsilon \text{ and } \|\mathbf{B} - \mathbf{J}\|_F^2 < \epsilon$$

end**Output:** $\mathbf{B}$ Now by fixing $\mathbf{D}$ and $\mathbf{B}$, we optimize $\mathbf{\Gamma}$ and $\mathbf{T}$ by employing the regularization (6.18) again:

$$\begin{aligned} \min_{\mathbf{\Gamma}, \mathbf{T}} \quad & \tau \|\mathbf{\Gamma}\|_F^2 + \delta \|\mathbf{T}\|_1 \\ \text{s.t.} \quad & \mathbf{A} - \mathbf{D} = \mathbf{B}\mathbf{\Gamma} + \mathbf{T}. \end{aligned} \tag{6.24}$$

Again, the ALM method is applied to covert (6.24) into an unconstrained minimization function

$$\begin{aligned} \tau \|\mathbf{\Gamma}\|_F^2 &+ \delta \|\mathbf{T}\|_1 + \text{tr}(\mathbf{Y}^T(\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} - \mathbf{T})) \\ &+ \frac{\mu}{2} \|\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} - \mathbf{T}\|_F^2 \end{aligned} \tag{6.25}$$

where $\mathbf{Y}$ is the Lagrange multiplier and μ a penalty parameter. The inexact ALM method to solve (6.25) is outlined in Algorithm 6. The same as Step 3 of Algorithm 5, the Step 2 of Algorithm 6 can be solved by soft-thresholding.

Algorithm 6: Solving Problem(6.24) by Inexact ALM

Input: $\mathbf{A}$, $\mathbf{D}$, $\mathbf{B}$, and parameter τ and δ .

Initialize: $\mathbf{T} = 0$, $\mathbf{Y} = 0$, $\mu = 10^{-3}$, $\mu_{max} = 10^6$, $\rho = 1.5$, and $\epsilon = 10^{-6}$.

while *not converged* **do**

1. fix the others and update $\mathbf{X}$ by

$$\mathbf{\Gamma} = \mu(2\tau\mathbf{I} + \mu\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{A} - \mathbf{D} - \mathbf{T} + \mathbf{Y}/\mu)$$

2. fix the others and update $\mathbf{T}$ by

$$\mathbf{T} = \arg \min_{\mathbf{T}} \frac{\delta}{\mu} \|\mathbf{T}\|_1 + \frac{1}{2} \|\mathbf{T} - (\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} + \mathbf{Y}/\mu)\|_F^2$$

3. update the multiplier

$$\mathbf{Y} = \mathbf{Y} + \mu(\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} - \mathbf{T})$$

4. update μ by $\mu = \min(\rho\mu, \mu_{max})$

5. check the convergence conditions:

$$\|\mathbf{A} - \mathbf{D} - \mathbf{B}\mathbf{\Gamma} - \mathbf{T}\|_F^2 < \epsilon$$

end

Output: $\mathbf{\Gamma}$

With the optimized $\mathbf{B}$ and $\mathbf{\Gamma}$, the remaining class-specific information in $\mathbf{A} - \mathbf{D}$ is separated from $\mathbf{B}\mathbf{\Gamma}$. The last step of the proposed dictionary decomposition is to transfer it to $\mathbf{D}$. This is achieved by using the regularization (6.18) again with fixed $\mathbf{B}$ and $\mathbf{\Gamma}$

$$\begin{aligned} \min_{\mathbf{D}, \mathbf{E}} \quad & \|\mathbf{D}\|_* + \eta \|\mathbf{E}\|_1 \\ \text{s.t.} \quad & \mathbf{A} - \mathbf{B}\mathbf{\Gamma} = \mathbf{D} + \mathbf{E}. \end{aligned} \tag{6.26}$$

As (6.26) is a standard low-rank matrix recovery problem, it can be solved by the algorithm in [117]. To ensure only random sparse noise being captured by $\mathbf{E}$, a high weight η should be assigned to $\|\mathbf{E}\|_1$.

The optimizations (6.21), (6.24) and (6.26) need iterate a few times due to two unknowns in the form of $\mathbf{B}\mathbf{\Gamma}$. We call the above proposed procedure the supervised low-rank (SLR) dictionary decomposition, which is summarized in Algorithm 7. SLR converges since the three alternative optimizations (6.21), (6.24) and (6.26) are all convex. Fig. 6.3 shows the convergence of SLR. The value of loss function (6.18) decays fast in the first a few iterations. Experiments show that only 3 to 5 iterations yield the stable best results. Thus, all experiments of this work iterate (6.21), (6.24) and (6.26) 4 times. SLR utilizes the low-rank matrix recovery to transfer information from the supervised assigned dictionary $\mathbf{B}$ to $\mathbf{D}$. The matrix $\mathbf{T}$ serves as a medium of the information transfer, which leads the final dictionary $\mathbf{D}$ containing more class-specific information than that used in the conventional linear discriminant analysis that is equivalent to the initial $\mathbf{D}$. The matrix $\mathbf{\Gamma}$ improves the quality of dictionary $\mathbf{B}$ for the representation (prediction) of unknowns in the proposed SDR framework. The proposed SLR dictionary decomposition not only delivers a class-specific dictionary $\mathbf{D}$ and a non-class-specific dictionary $\mathbf{B}$, it also alleviates the other fundamental problem of SRC – corrupted training data, since the random sparse noise of training data $\mathbf{E}$ is removed from the two learnt dictionaries.

Algorithm 7: SLR Dictionary Decomposition

Input: $\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_i, \dots, \mathbf{A}_c]$ and parameter κ .

Initialize: $\mathbf{\Gamma} = \mathbf{I}$, $\mathbf{A}_i = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i^T$, $\mathbf{D}_i = \mathbf{U}_i(1:m, 1) \mathbf{\Sigma}_i(1, 1) \mathbf{V}_i(1:n_i, 1)^T$,

$\mathbf{D} = [\mathbf{D}_1, \dots, \mathbf{D}_i, \dots, \mathbf{D}_c]$.

for $l:\kappa$ **do**

1. obtain $\mathbf{B}$ by solving the optimization problem (6.21) using Algorithm 5.
2. update $\mathbf{\Gamma}$ by solving the optimization problem (6.24) using Algorithm 6.
3. update $\mathbf{D}$ by solving standard low-rank matrix recovery problem (6.26).

end

Output: $\mathbf{D}$ and $\mathbf{B}$

Fig. 6.4 visualizes some results of the proposed SLR dictionary decomposition on AR database. Training images of one of 100 persons corrupted by strong sparse noise in $\mathbf{A}$ are shown in Fig. 6.4 (a). Their initial assignments to $\mathbf{D}$ and $\mathbf{B}$ are shown in Fig. 6.4 (b) and (c), respectively. The final components of Fig. 6.4 (a) in $\mathbf{A}$ decomposed by the proposed SLR in

Figure 6.3: An example of SLR's convergence on the AR database.

Figure 6.4: Sample images produced by the proposed SLR: images of a person in A corrupted by random noise (a); their initial components assigned to D (b) and B (c); final images in D (d), $B\Gamma$ (e) and E (f).

Figure 6.5: Examples of RPCA and the proposed SLR: training images (a); low-rank (b) and sparse (c) images from RPCA; class-specific (d), non-class-specific (e) and sparse (f) images decomposed by SLR.

$\mathbf{D}$, $\mathbf{B}\Gamma$ and $\mathbf{E}$ are shown in Fig. 6.4 (d), (e) and (f), respectively. Fig. 6.4 also serves as a visual comparison of the dictionaries decomposed by the proposed SLR to those used in SSRC [95], a further development from ESRC [94]. SSRC uses the class-wise means of $\mathbf{A}$ as one dictionary and the class-wise centralized $\mathbf{A}$ as the other. They are similar to Fig. 6.4 (b) and (c) except for a scaling factor.

Fig. 6.5 shows a further visual example of the proposed SLR comparing to RPCA. The data set contains 301 face images from AR database, where 1 wears scarf and each of 100 persons has 1 image wearing sunglasses and 2 undisguised images. Fig. 6.5 (b) and (c) show that sunglasses and scarf in Fig. 6.5 (a) are not fully removed from the dictionary but shared by it and the sparse noise by RPCA. In contrast, they are fully removed from the class-specific dictionary by SLR as shown in Fig. 6.5 (d). Moreover, sunglasses, though much smaller than scarf, is allotted in the non-class-specific dictionary. This is desirable as over 30% samples in the data set wear sunglasses and so is expected for the query images. Allotting it in $\mathbf{B}$ increases its representation power. This improves rather than damages the classification.

6.4 Experiments

The proposed SDR framework with the proposed SLR dictionary decomposition algorithm, SDR-SLR, is evaluated on 4 face databases: CMU Multi-PIE [164], Extended Yale B [143], AR [139] and FERET [174]. It is compared with related state-of-the-art SRC [67], ESRC [94], SSRC [95], LR (Low-Rank SRC) and LRSI (LR with structural incoherence) [114] as well as baseline classifiers NN (nearest neighbor) and LRC [71]. The parameters of the proposed SDR-SLR approach are fixed over all experiments of this chapter to $\beta = \sigma = 10$ for the SDR and $\kappa = 4$, $\lambda = 1$, $\tau = 0.01$, $\delta = 0.8\nu$ and $\eta = 1.2\nu$ for SLR, where $\nu = 1/\sqrt{\max(\text{size}(\mathbf{D}))}$ [147], though better performance may be achieved if they are fine tuned to fit each specific experiment.

The CMU Multi-PIE database contains face images captured in 4 sessions with variations in illumination, expression and pose. The first 105 subjects that appear in all 4 sessions are used in the experiments. Images are cropped based on the eye locations provided by [145] and down-sampled to 50×40 pixels.

The cropped Extended Yale B database has 2,414 frontal-face images of size 192×168 from 38 subjects, captured under different lighting conditions. 64 images per subject are divided into 5 subsets according to the angle of the light direction with 7, 12, 12, 14 and 19 images respectively in subset 1 to 5.

The AR database used in this chapter has 2600 frontal-face images of size 165×120 from 100 subjects (50 males and 50 females). In each of two separate sessions, 7 undisguised images with expression or illumination variation, 3 images in sunglasses and 3 images in scarf disguise are taken from each subject.

A subset of FERET database contains 1024 images from 256 subjects with 4 frontal images per subject. Images are normalized by an affine transformation, scaled and cropped to the size of 121×121 . Fig. 6.6 shows the normalized and cropped images.

Figure 6.6: Normalized images from FERET database.

6.4.1 Face Recognition on Single Variation

This subsection tests the effectiveness of the proposed SDR-SLR on isolated single type of variation: illumination, expression or pose. Misalignment is not tested as it can be alleviated by using alignment insensitive features such as histogram of gradient [175] or LBP [176, 177]. Recent efforts that recover the misalignment can be found in [149, 178, 179].

Experiment 1(illumination): following the procedure in [164], 18 flash-only images are generated as the difference between flash images (illuminations $\{1-18\}$) and non-flash image (illumination $\{0\}$) of the Multi-PIE database as shown in Fig. 6.7(a). For each of these 18 different illuminations, 4 frontal images with neutral expression per subject are chosen from 4 different sessions, which produces $18 \times 4 \times 105 = 7560$ images for training and testing. For each subject, t different illuminations are randomly selected for training and the rest $18 - t$ illuminations are used for testing. The averaged recognition rates of 10 runs are plotted against the reducing t in Fig. 6.8. Much to our surprise, ESRC, SSRC, LR and LRSI, which try to solve the problems of SRC, in fact underperform SRC. Nevertheless, the proposed SDR-SLR consistently outperforms the others in all cases where the performance gain increases with decreasing illumination variations in the training data.

Experiment 2(illumination): to further verify the performances of various methods in tackling the illumination variation, another experiment is done on Extended Yale B database. For each subject, 2 subsets are randomly chosen for training and the other 3 subsets for testing. Different image sizes are tested. The average recognition rates of 10 runs in Table 6.1 further show that ESRC, SSRC, LR and LRSI underperform SRC on data set that has only illumination variation. Again, the proposed SDR-SLR visibly outperforms the others consistently for all image sizes.

Experiment 3(expression): all frontal images with illumination $\{7\}$ are taken from the

Figure 6.7: Sample images in Multi-PIE database with variations of (a) illumination, (b) expression, (c) pose.

Table 6.1: Recognition Rate on Extended Yale B

dimensions	810	1400	2016
NN	42.1%	43.9%	44.6%
LRC [71]	72.8%	73.1%	73.2%
SRC [67]	<u>77.9%</u>	<u>78.5%</u>	<u>78.7%</u>
ESRC [94]	77.6%	<u>78.5%</u>	78.6%
SSRC [95]	74.8%	75.2%	76.5%
LR [114]	77.3%	77.6%	78.0%
LRSI [114]	77.8%	78.0%	78.5%
SDR-SLR	<u>81.7%</u>	<u>84.3%</u>	<u>84.5%</u>

Figure 6.8: Recognition rate versus number of training illuminations per subject on Multi-PIE database.

Figure 6.9: Recognition rate versus number of training expressions per subject on Multi-PIE database.

Table 6.2: Recognition Rate on the Multi-PIE Face Database with Pose Variation.

t	4	3	2
NN	17.0%	13.7%	11.4%
LRC [71]	41.9%	29.5%	18.4%
SRC [67]	65.8%	57.8%	45.2%
ESRC [94]	73.0%	69.0%	57.3%
SSRC [95]	<u>79.3%</u>	<u>73.8%</u>	<u>61.2%</u>
LR [114]	66.2%	59.4%	46.6%
LRSI [114]	66.6%	58.6%	45.2%
SDR-SLR	<u>87.2%</u>	<u>81.4%</u>	<u>67.1%</u>

Multi-PIE database. The 6 expressions (neutral, smile, surprise, squint, disgust and scream) are shown in Fig. 6.7(b). For each subject, t different expressions are randomly selected for training and the rest $6 - t$ expressions are used for testing. The averaged recognition rates of 10 runs are plotted against the reducing t in Fig. 6.9. It shows that LR and LRSI performs similar or even inferior to SRC as the database is uncorrupted. As no similar expression of the query image appears in the training samples of the same subject, ESRC, SSRC and SDR-SLR significantly outperform SRC. The proposed SDR-SLR consistently and visibly performs the best for all values of t where the performance gain increases with decreasing expression variations in the training data.

Experiment 4(pose): 20 images of neutral expression from 4 sessions captured by 5 cameras (from -30% to +30%) and flashed by the capturing camera are taken from Multi-PIE database as shown in Fig. 6.7(c). For each subject, t different poses are randomly chosen for training and the rest $5 - t$ poses are used for testing. Table 6.2 details the average results over 10 runs. LR and LRSI only slightly outperform SRC on the uncorrupted data. Similar to the experiment 3, ESRC, SSRC and SDR-SLR significantly outperform SRC, where the proposed SDR-SLR consistently and visibly performs the best for all values of t .

6.4.2 Face Recognition on Mixed Variations

This subsection tests the effectiveness of the proposed SDR-SLR on databases whose face images have mixed types of variations without image corruption.

Experiment 5: all undisguised images from Session 1 of the AR database are used for training and those from Session 2 are used for testing. Images down-sampled to 4 different sizes are tested. The results are recorded in Table 6.3. Unfortunately, ESRC, SSRC, LR and LRSI again do not outperform SRC though the data have the both expression and illumination variations. Probably the training and testing sets, though taken from two separate sessions, have very similar variations. Nevertheless, the proposed SDR-SLR approach again visibly outperforms the others consistently for all image sizes.

Table 6.3: Recognition Rate on All Undisguised Images of AR Dataset

dimensions	540	850	1200	2200
NN	69.5%	70.4%	71.2%	71.8%
LRC [71]	74.1%	75.2%	76.0%	76.4%
SRC [67]	90.7%	<u>91.6%</u>	<u>92.4%</u>	<u>92.8%</u>
ESRC [94]	90.9%	90.8%	91.4%	91.8%
SSRC [95]	87.7%	90.7%	90.7%	90.7%
LR [114]	90.8%	91.3%	91.7%	<u>92.8%</u>
LRSI [114]	<u>91.1%</u>	91.3%	91.6%	92.5%
SDR-SLR	<u>96.4%</u>	<u>96.7%</u>	<u>96.9%</u>	<u>97.6%</u>

The classification running time of each approach is evaluated under the Matlab programming environment and on a desktop of 3.5GHZ with 16G RAM. For dimension equal to 2200, it is 0.03 for LRC, 0.27 for SRC, 0.78s for ESRC, 0.59s for SSRC, 0.08s for LR and LRSI, 2.77s for SDR.

Experiment 6: to show the cases where the problems of SRC become severe, we repeat the above experiment for the image dimensions of 2200 with reduced number of training samples

per subject to 6, 5, 4, and 3. The average results of 10 runs in Table 6.4 show that, except for the case of 6 training samples, LR and LRSI underperform SRC as the training data have no corruption. ESRC and SSRC start outperforming SRC from 5 training samples, yet marginally. The best accuracy gain of ESRC or SSRC over SRC is 2.8% at 3 training samples. In contrast, that of the proposed SDR-SLR reaches almost 10%.

Table 6.4: Recognition Rate of Fewer Training Samples of AR Dataset

train. samples	6	5	4	3
NN	65.4%	60.4%	53.5%	46.1%
LRC [71]	72.8%	67.4%	60.5%	53.7%
SRC [67]	91.7%	88.7%	87.7%	82.3%
ESRC [94]	90.4%	<u>89.4%</u>	88.4%	<u>85.1%</u>
SSRC [95]	90.3%	<u>89.4%</u>	<u>89.0%</u>	84.7%
LR [114]	<u>91.9%</u>	87.8%	86.7%	79.5%
LRSI [114]	91.7%	88.6%	87.3%	79.7%
SDR-SLR	<u>96.4%</u>	<u>96.3%</u>	<u>95.6%</u>	<u>92.1%</u>

Experiment 7: for each subject of the FERET data set, we randomly pick out two images, one for training and the other for testing. Then, the half of the remaining 512 images is randomly taken for training and the rest are for testing. Thus, there are 512 images in training and the other 512 images in testing with at least one in training and another in testing per subject. Images down-sampled to 4 different sizes are tested. The average results of 10 runs in Table 6.5 show that LR and LRSI have about the same accuracy as SRC. For this small number of training samples per class, the both ESRC and SSRC outperform SRC with the gains between 2% and 4.1%. Much more significant gains in recognition performance over SRC are achieved by the proposed SDR-SLR consistently over all different image dimensions. They are between 9.2% and 10.7%.

Table 6.5: Recognition Rate on FERET Database

dimensions	400	900	1600	2500
NN	68.5%	67.6%	67.0%	66.7%
LRC [71]	68.0%	67.5%	66.9%	66.5%
SRC [67]	80.8%	80.1%	79.9%	79.1%
ESRC [94]	<u>83.5%</u>	83.5%	<u>82.9%</u>	82.4%
SSRC [95]	83.3%	<u>84.2%</u>	81.9%	<u>82.9%</u>
LR [114]	80.9%	79.6%	79.2%	79.0%
LRSI [114]	81.1%	80.0%	79.7%	79.4%
SDR-SLR	<u>90.0%</u>	<u>90.1%</u>	<u>90.6%</u>	<u>89.5%</u>

6.4.3 Face Recognition on Corrupted Data

This subsection tests the effectiveness of various approaches on training data with different types and levels of corruption.

Experiment 8: the undisguised subset of AR database is used. 5 images per subject from Session 1 are randomly chosen as training set and all images from Session 2 are used as test set. All images are resized to 50×44 pixels. For each training image, a certain percentage of its pixels are randomly replaced by noise uniformly distributed between the minimal and the maximal pixel value. The average recognition rate over 10 runs is plotted against the noise level in Fig. 6.10. It shows that ESRC and SSRC perform about the same as SRC. LR and LRSI slightly outperform SRC thanks to their low-rank recovery of the training data. Fig. 6.10 demonstrates that the proposed SDR-SLR significantly outperforms all other algorithms consistently for all levels of corruption.

Experiment 9: besides the random pixel corruption, we further test different approaches in coping with random block occlusion on Extended Yale B database. Same as [67], Subsets 1 and 2 containing 719 images are taken for training and Subset 3 containing 455 images is used for testing. Images are resized to 48×42 . Also same as [67], different portions of images, from

Figure 6.10: Recognition rate versus percentage of uniform noise on AR database.

Figure 6.11: Examples of images of Extended Yale B database with different levels of block occlusion. From left to right, 20%, 30%, 40% and 50% of images are occluded, respectively.

20% to 50%, are occluded by an unrelated image at random locations as shown in Fig. 6.11. But different from [67] that corrupts all test images, we corrupt half of the training and half of the testing images. The result over 10 runs is plotted against the level of occlusion in Fig. 6.12.

It seems a surprise that for this corrupted data, LR and LRSI do not consistently have better performance than SRC but ESRC and SSRC do. Possible reasons could be that the same image is used to occlude all 50% training and testing images so that the block occlusion causes many training samples having a same occluded area and even same as many testing images. As a result, the occlusion appears more like a common variation than the random corruption. Nevertheless, the proposed SDR-SLR outperforms all other algorithms consistently for all levels of occlusion.

6.4.4 Face Recognition with Real Disguised Images

This subsection tests the effectiveness of various approaches in dealing with real possible malicious occlusions in both training and testing samples. Using all images of size 55×40 in AR database,

Experiment 10 considers the following 4 scenarios:

Figure 6.12: Recognition rate versus percentage of occlusion on extended Yale B database.

Sunglasses only: for each of the 100 subjects, all 7 undisguised images and 1 image with sunglasses (random chosen) from Session 1 are selected as training set. All remaining undisguised images and images with sunglasses are used for testing. So, there are 8 training and 12 testing images per subject.

Scarf only: replace images with sunglasses in the above scenario by images with scarf.

Mixed 1 (Sunglasses and Scarf): for each subject, all undisguised images, 1 image with sunglasses (random chosen) and 1 image with scarf (random chosen) from Session 1 are selected as training set. All remaining images are used for testing. So, there are 9 training and 17 testing images per subject.

Mixed 2 (Sunglasses or Scarf): 1 image randomly taken from the 6 disguised images per subject and all undisguised images from Session 1 are used for training. All remaining images are used for testing. So, there are 8 training and 18 testing images per subject.

The first 3 scenarios are exactly the same as those used in [95, 114]. The forth scenario is new and more challenge because every class has two different disguise types in its testing data, one of which, though presents in some other classes, is absent in its training data. We repeat each scenario three times and the average results are recorded in Table 6.6. It shows that LR or LRSI is consistently but very marginally better than SRC. The both ESRC and SSRC outperform SRC more visibly, where SSRC is consistently the second best performer, which gains accuracy over SRC between 1.1% and 6.6%. The proposed SDR-SLR consistently performs the best and its accuracy gain over SRC ranges from 7.5% to 13.1%.

Table 6.6: Recognition Rate on the AR Database.

Scenario	Sunglass	Scarf	Mixed 1	Mixed 2
NN	51.4%	49.0%	41.8%	35.8%
LRC [71]	69.8%	64.8%	64.2%	47.3%
SRC [67]	88.8%	86.6%	85.8%	79.9%
ESRC [94]	89.6%	89.5%	88.9%	84.5%
SSRC [95]	<u>89.9%</u>	<u>90.1%</u>	<u>89.3%</u>	<u>86.5%</u>
LR [114]	88.6%	86.8%	86.1%	79.9%
LRSI [114]	89.0%	86.7%	86.3%	80.2%
SDR-SLR	<u>96.3%</u>	<u>94.5%</u>	<u>95.3%</u>	<u>93.0%</u>

6.5 Summary

Sparse representation shows some merits in holistic image classification. In seeking a sparse representation of a query image, every sample competes against the others to gain its share. This well matches the general classification objective that only one class should stand out from the rest. However, all information in images participates in this competition. There is no reason why the significant coefficients must coincide with a specific one of many different forms of class grouping, especially for some applications where the class-specific information takes up only a very small portion of that in images. As a result, SRC requires sufficient representative samples for every class, with which the non-class-specific information can be nulled out. This requirement may not be fulfilled in many computer vision problems such as face recognition. Another problem of SRC is the corrupted training data though it is robust to the corrupted query image thanks to the ℓ_1 minimization of the error.

This work alleviates the two fundamental problems of SRC by a sparse- and dense-hybrid representation (SDR) based on a supervised low-rank (SLR) dictionary decomposition/learning. In the proposed SDR framework, every sample only uses its class-specific component to compete against the others collaboratively with the non-class-specific component of all samples. This makes the sparse code of the proposed SDR completely coincide with the specifically as-

signed class membership. The co-representation by the non-class-specific component largely relaxes the requirement of representative samples for every class. The class-specific dictionary decomposed by the proposed SLR captures more information than that used in the linear discriminant analysis. The sparse outlier pixels and occlusions of the training data are also separated from the two decomposed low-rank dictionaries. This alleviates the second problem of SRC with corrupted training data.

Extensive experiments on 4 face image databases have verified the effectiveness and advancement of the proposed SDR-SLR approach. It consistently and visibly outperforms SRC and its related extensions in all experiments, whether the training data are corrupted or not and whether every class has sufficient representative training samples or not. In case the training data are corrupted or are not representative for some classes, the performance gains of the proposed SDR-SLR approach are significant.

Chapter 7

Conclusions and Future Work

In this thesis, we analyze the merits and limitations of the face recognition methods built on sparse representation. To address the problems of existing methods, we propose several sparse representation based approaches for face recognition.

Recently, SRC has been applied to face recognition and significantly differentiated itself from the traditional classifiers in three aspects. One is the utilization of training samples of all classes collaboratively to represent the query images and another is the sparse representation code that coincides with the general classification target. The last is the ℓ_1 -norm minimization of the representation error that enables SRC to recognize query images sparsely corrupted by outlier pixels and occlusions. Thanks to these merits, SRC achieves some impressive results, which lead to many interesting extensions in classification, feature extraction, dictionary learning, and face alignment.

In spite of its success, SRC has some limitations in face recognition application. As SRC introduces an identity matrix as occlusion dictionary, it is with huge computational burden. Therefore, some works focus on accelerating the runtime of sparse representation procedure. In addition, the sparse assumption of noise limits the application of SRC. To tackle this problem, a number of frameworks are introduced to handle the non-sparse corruption by utilizing prior

knowledge (e.g., spatial continuity of the occlusion). In order to achieve a more discriminative representation for recognition task, many works replace the sparse constraint of coefficients with other restrictions, such as locality and group-wise sparse constraint. Instead of directly using the training data as dictionary in SRC, which may not be discriminative for classification, some approaches are developed to learn the optimum dictionaries by considering both reconstruction and discrimination power.

Although numerous sparse representation based methods are proposed in recent years, some problems of sparse representation are not yet well addressed. How to appropriately learn the underlying structure of labelled training data, which captures both sparse and discriminative information for feature extraction? How to handle the severely corrupted query image? How to achieve a promising result when sufficient representative and well controlled training data can not be guaranteed? To address these problems, we propose several sparse representation based approaches.

This chapter first concludes the contributions of thesis in Section 7.1, which is followed by some potential directions of future work in Section 7.2.

7.1 Contributions

First, we propose a supervised sparse representation (SSR) and a discriminative sparse preserving embedding (DSPE) for feature extraction. Very recently, a sparsity preserving method is introduced for feature extraction. As an unsupervised approach, sparse representation may not learn sufficient discrimination for dimension reduction. By utilizing of the crucial label information, the proposed SSR learns a discriminative representation. Furthermore, the proposed DSPE captures the discriminative sparse structure of labelled training data based on SSR, and finds a low dimensional subspace that reduces the within class distances and keeps the between

class distances. Compared with related state-of-the-art approaches, the extensive experiments show that the proposed DSPE consistently achieves the best recognition accuracies on benchmark face databases.

Second, to address the limitation of SRC in handling the severely occluded query face, a modular weighted global sparse representation (WGSR) is proposed for robust face recognition. An image is first divided into modules and each module is processed separately to determine its reliability. A reconstructed image from the modules weighted by their reliability is formed for the robust recognition. We proposed to use the modular sparsity and residual jointly to determine the modular reliability. The proposed framework advances both the modular and global sparse representation approaches, particularly in dealing with disguise, large illumination variations and expression changes.

As sparse representation is unsupervised, the crucial label information of training data is not used in searching the coefficients. To alleviate the problem of unsupervised optimization of SRC, we propose a class-wise sparsity representation (CSR) by employing the class label information. The proposed CSR seeks an optimum representation of the query image such that it simultaneously optimizes the reconstruction error and class-wise sparsity. To tackle the problem of insufficient training data, we further propose a collaborative patch framework, together with CSR, named CSR-CP. Different from the patch based methods, which optimize each patch independently, the proposed CSR-CP reconstructs the whole query sample by sub-images of all subjects across all patches within one framework and searches a patch-wise and class-wise sparse representation. Extensive experiments on several benchmark databases verify the merits and significance of the proposed CSR-CP approach.

With sufficient representative and carefully controlled training data, SRC achieves some impressive results. A violation of the two assumptions leads to a poor performance of SRC. Aiming at tackling these limitations, we propose a sparse- and dense-hybrid representation

(SDR) framework with supervised low-rank (SLR) dictionary decomposition. The query image can be well represented by a sparse combination of the class-specific dictionary and a dense combination of non-class-specific dictionary. We further propose a procedure of supervised low-rank dictionary decomposition to facilitate the proposed SDR framework. In addition, the problem of the corrupted training data is also alleviated by the proposed SLR dictionary decomposition. The proposed SDR-SLR consistently and visibly outperforms SRC and its related extensions in all conducted experiments, especially when the training data are corrupted or not representative for some classes.

7.2 Future Work

In this section, some possible future research directions are presented.

Although sparse representation achieves impressive results in computer vision compared to the traditional dense representation (i.e., ℓ_2 -norm regularization), it requires a huge computational burden due to the iterative optimization procedure. As a result, sparse representation is difficult to be applied to some real-time and mobile phone based applications. How to investigate a fast algorithm meanwhile keeping or even enhancing the performances is an interesting future direction.

We also want to extend the proposed work to more general face analysis (e.g., expression, age and gender recognition). Take expression as an example. To recognize the facial expression, which is one of the most natural and immediate ways for human to express emotions, is an interesting and challenging problem. By utilizing the proposed WGSF, the most discriminative modules for expression analysis can be retained for classification.

In this thesis, our works are limited to recognize face with only one single query image. In

some realistic scenarios, a set query images are available, which can be taken under different variations. As there are more discriminative information, the performances of query set based approach should be much better than those of single query image based one. We will explore how to appropriately represent and measure the distance between sets (i.e., query and training sets) in the future work.

Compared with the still image, video can capture more abundant information, which is an important reason why video sharing websites (e.g. Youtube) are with increasing popularity. It would be very useful to sort the videos or segment a long video into several small ones based on the identities of the captured subjects. Therefore, a fast and accurate system to recognize a face based on the input video is urgent. Compared with still image, video frames can be more noisy and with lower resolution. In addition, the data size of video is significant large. How to tackle these problems and develop a reliable face recognition approach based on video could be one of our future directions.

Author's Publications

Journal Papers

1. **Jian Lai** and Xudong Jiang, "Modular Weighted Global Sparse Representation for Robust Face Recognition", *IEEE Signal Processing letters*, vol. 19, no. 9, pp. 571-574, September 2012.
2. Xudong Jiang and **Jian Lai**, "Sparse And Dense Hybrid Representation via Dictionary Decomposition for Face Recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1067-1079, May, 2015.
3. **Jian Lai** and Xudong Jiang, "Class-wise Sparse and Collaborative Patch Representation for Robust Face Recognition", *IEEE Trans. Image Processing*, under major revision.

Conference Papers

1. **Jian Lai** and Xudong Jiang, "Robust Face Recognition Using Trimmed Linear Regression", in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.
2. **Jian Lai** and Xudong Jiang, "Discriminative Sparsity Preserving Embedding for Face Recognition", in *IEEE International Conference Image Processing (ICIP)*, 2013.
3. **Jian Lai** and Xudong Jiang, "Supervised Trace Lasso for Robust Face Recognition", *IEEE International Conference Multimedia and Expo (ICME)*, 2014.

Bibliography

- [1] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 103–108, jan 1990.
- [2] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: a survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65 – 77, 1992.
- [3] R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, pp. 705 –741, may 1995.
- [4] Y. Adini, Y. Moses, and S.Ullman, "Face recognition: the problem of compensating for changes in illumination direction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 721–732, Jul 1997.
- [5] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *IEEE Trans. Neural Networks*, vol. 8, pp. 98–113, Jan 1997.
- [6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 711 –720, July 1997.
- [7] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, p-p. 775–779, Jul 1997.
- [8] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 23–38, Jan 1998.
- [9] P. J. Phillips, "Support vector machines applied to face recognition," in *Advances in Neural Information Processing Systems*, pp. 803–809, 1999.
- [10] C. Liu and H. Wechsler, "Evolutionary pursuit and its application to face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 570–582, Jun 2000.

- [11] C. Liu and H. Wechsler, "A shape- and texture-based enhanced fisher classifier for face recognition," *IEEE Trans. Image Processing*, vol. 10, pp. 598–608, Apr 2001.
- [12] J. Beveridge, K. She, B. Draper, and G. Givens, "A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. I-535–I-542 vol.1, 2001.
- [13] A. S. Georgiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 643–660, June 2001.
- [14] M. J. Er, S. Q. Wu, J. Lu, and H. L. Toh, "Face recognition with radial basis function (rbf) neural networks," *IEEE Trans. Neural Networks*, vol. 13, pp. 697–710, May 2002.
- [15] Y. Gao and M. Leung, "Face recognition using line edge map," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 764–779, Jun 2002.
- [16] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, pp. 399–458, Dec. 2003.
- [17] J. Yang, D. Zhang, A. Frangi, and J. Y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, pp. 131–137, Jan 2004.
- [18] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, pp. 2037–2041, Dec 2006.
- [19] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, pp. 4–20, Jan 2004.
- [20] A. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *IEEE Trans. Information Forensics and Security*, vol. 1, pp. 125–143, June 2006.
- [21] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Springer, 2008.
- [22] J. Daugman, "How iris recognition works," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, pp. 21–30, Jan 2004.
- [23] R. Wildes, "Iris recognition: an emerging biometric technology," *Proceedings of the IEEE*, vol. 85, pp. 1348–1363, Sep 1997.

- [24] L. Ma, T. N. Tan, Y. H. Wang, and D. X. Zhang, "Efficient iris recognition by characterizing key local variations," *IEEE Trans. Image Process.*, vol. 13, pp. 739–750, June 2004.
- [25] J. Daugman, "The importance of being random: statistical principles of iris recognition," *Pattern Recognition*, vol. 36, no. 2, pp. 279 – 291, 2003.
- [26] A. K. Hrechak and J. A. McHugh, "Automated fingerprint recognition using structural matching," *Pattern Recognition*, vol. 23, no. 8, pp. 893 – 904, 1990.
- [27] X. D. Jiang and W. Y. Yau, "Fingerprint minutiae matching based on the local and global structures," in *Int'l. Conf. Pattern Recognition*, vol. 2, pp. 1038–1041 vol.2, 2000.
- [28] X. D. Jiang, W. Y. Yau, and W. Ser, "Detecting the fingerprint minutiae by adaptive tracing the gray-level ridge," *Pattern Recognition*, vol. 34, no. 5, pp. 999 – 1013, 2001.
- [29] X. D. Jiang and W. Ser, "Online fingerprint template improvement," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 1121–1126, Aug 2002.
- [30] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*. Springer Publishing Company, Incorporated, 2nd ed., 2009.
- [31] J. Zhang, R. Jin, Y. Yang, and A. G. Hauptmann, "Modified logistic regression: An approximation to svm and its applications in large-scale text categorization," in *Proc. IEEE Int'l. Conf. Machine Learning*, Press, 2003.
- [32] A. Kong, D. Zhang, and M. Kamel, "A survey of palmprint recognition," *Pattern Recognition*, vol. 42, no. 7, pp. 1408 – 1418, 2009.
- [33] H. Chen and B. Bhanu, "Human ear recognition in 3d," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, pp. 718–737, April 2007.
- [34] B. Bhanu and H. Chen, *Human Ear Recognition by Computer*. Springer Publishing Company, Incorporated, 1 ed., 2008.
- [35] L. Wang, T. N. Tan, H. Z. Ning, and W. M. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, pp. 1505–1518, Dec 2003.
- [36] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, pp. 316–322, Feb 2006.
- [37] D. C. Tao, X. L. Li, X. D. Wu, and S. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, pp. 1700–1715, Oct 2007.

- [38] D. L. Donoho and M. Elad, “Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization,” *Proc. Natl Acad. Sci. USA*, pp. 2197–2202, Mar. 2003.
- [39] E. J. Candès, J. K. 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.
- [40] E. Candes and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?,” *IEEE Trans. Information Theory*, vol. 52, pp. 5406–5425, dec. 2006.
- [41] D. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, “Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit,” *IEEE Trans. Information Theory*, vol. 58, pp. 1094–1121, Feb 2012.
- [42] D. Donoho, “Compressed sensing,” *IEEE Trans. Information Theory*, vol. 52, pp. 1289–1306, April 2006.
- [43] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.
- [44] E. Candes, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” *IEEE Trans. Information Theory*, vol. 52, pp. 489–509, Feb 2006.
- [45] Y. Pati, R. Rezaeiifar, and P. Krishnaprasad, “Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition,” in *Proc. 27th Annu. Asilomar Conf. Signals, Systems and Computers*, pp. 40–44 vol.1, 1993.
- [46] J. Tropp and A. Gilbert, “Signal recovery from random measurements via orthogonal matching pursuit,” *IEEE Trans. Information Theory*, vol. 53, pp. 4655–4666, Dec 2007.
- [47] D. Needell and R. Vershynin, “Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit,” *Foundations of Computational Mathematics*, vol. 9, no. 3, pp. 317–334, 2009.
- [48] D. Needell and R. Vershynin, “Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 310–316, April 2010.
- [49] H. Huang and A. Makur, “Backtracking-based matching pursuit method for sparse signal reconstruction,” *IEEE Signal Processing Letters*, vol. 18, pp. 391–394, July 2011.

- [50] J. C. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.
- [51] J. C. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Processing*, vol. 19, pp. 2861–2873, Nov 2010.
- [52] S. Y. Yang, M. Wang, Y. Chen, and Y. Sun, "Single-image super-resolution reconstruction via learned geometric dictionaries and clustered sparse coding," *IEEE Trans. Image Processing*, vol. 21, pp. 4016–4028, Sept 2012.
- [53] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2790–2797, June 2009.
- [54] E. Elhamifar and R. Vidal, "Clustering disjoint subspaces via sparse representation," in *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing*, pp. 1926–1929, March 2010.
- [55] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, pp. 2765–2781, Nov 2013.
- [56] J. F. Cai, H. Ji, C. Q. Liu, and Z. W. Shen, "Blind motion deblurring from a single image using sparse approximation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 104–111, June 2009.
- [57] G. S. Yu, G. Sapiro, and S. Mallat, "Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity," *IEEE Trans. Image Processing*, vol. 21, pp. 2481–2499, May 2012.
- [58] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [59] M. Elad and M. Aharon, "Image denoising via learned dictionaries and sparse representation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 895–900, June 2006.
- [60] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, 2008.
- [61] X. Mei and L. H. B., "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 33, pp. 2259–2272, Nov. 2011.

- [62] Z. J. Han, J. B. Jiao, B. C. Zhang, Q. X. Ye, and J. Z. Liu, "Visual object tracking via sample-based adaptive sparse representation (adasr)," *Pattern Recognition*, vol. 44, no. 9, pp. 2170 – 2183, 2011.
- [63] T. Z. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2042–2049, June 2012.
- [64] S. P. Zhang, H. X. Yao, X. Sun, and X. S. Lu, "Sparse coding based visual tracking: Review and experimental comparison," *Pattern Recognition*, vol. 46, no. 7, pp. 1772 – 1788, 2013.
- [65] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, and R. C. R. G. Baraniuk, "Compressive sensing for background subtraction," in *Proc. European Conf. Computer Vision*, pp. 155–168, 2008.
- [66] M. Dikmen and T. Huang, "Robust estimation of foreground in surveillance videos by sparse error estimation," in *Proc. Int'l Conf. Pattern Recognition*, pp. 1–4, Dec 2008.
- [67] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, pp. 210–227, Feb. 2009.
- [68] S. Z. Li and J. W. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. Neural Networks*, vol. 10, pp. 439 –443, Mar. 1999.
- [69] J. T. Chien and C. C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 1644 – 1649, Dec. 2002.
- [70] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 2, pp. 218–233, 2003.
- [71] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, pp. 2106–2112, Nov. 2010.
- [72] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 84–91, 1994.
- [73] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Information Theory*, vol. 51, pp. 4203 – 4215, dec. 2005.

- [74] D. L. Donoho, “For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
- [75] G. Davis, S. Mallat, and M. Avellaneda, “Adaptive greedy approximations,” *Constructive Approximation*, vol. 13, no. 1, pp. 57–98, 1997.
- [76] M. Osborne, B. Presnell, and B. Turlach, “A new approach to variable selection in least squares problems,” *IMA Journal of Numerical Analysis*, vol. 20, no. 3, pp. 389–403, 2000.
- [77] I. Daubechies, M. Defrise, and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [78] D. Malioutov, M. Cetin, and A. Willsky, “Homotopy continuation for sparse signal representation,” in *Proc. IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. v/733 – v/736 Vol. 5, march 2005.
- [79] P. Combettes and V. Wajs, “Signal recovery by proximal forward-backward splitting,” *SIAM Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.
- [80] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” *Ann. Statist.*, vol. 32, pp. 407–499, 04 2004.
- [81] M. Figueiredo, R. Nowak, and S. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, pp. 586 –597, dec. 2007.
- [82] K. Koh, S. Kim, and S. Boyd, “An interior-point method for large-scale ℓ_1 -regularized logistic regression,” *J. Mach. Learn. Res.*, vol. 8, pp. 1519–1555, Dec. 2007.
- [83] E. Hale, W. Yin, and Y. Zhang, “A fixed-point continuation method for ℓ_1 -regularized minimization with applications to compressed sensing,” *Technical Report CAAM Technical Report TR07-07, Rice University, Houston, TX*, 2007.
- [84] D. Donoho and Y. Tsaig, “Fast solution of ℓ_1 -norm minimization problems when the solution may be sparse,” *IEEE Trans. Information Theory*, vol. 54, pp. 4789–4812, Nov 2008.
- [85] E. Hale, W. Yin, and Y. Zhang, “Fixed-point continuation for ℓ_1 -minimization: Methodology and convergence,” *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1107–1130, 2008.

- [86] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [87] S. Becker, J. Bobin, and E. Cands, "Nesta: A fast and accurate first-order method for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011.
- [88] A. Yang, Z. Zhou, A. Balasubramanian, S. Sastry, and Y. Ma, "Fast ℓ_1 -minimization algorithms for robust face recognition," *IEEE Trans. Image Processing*, vol. 22, no. 8, pp. 3234–3246, 2013.
- [89] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *Proc. European Conf. Computer Vision*, vol. 6316, pp. 448–461, 2010.
- [90] R. He, W. S. Zheng, and B. G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 33, pp. 1561–1576, Aug. 2011.
- [91] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, pp. 1753–1766, May 2013.
- [92] W. Liu, P. Pokharel, and J. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Trans. Signal Processing*, vol. 55, pp. 5286–5298, nov. 2007.
- [93] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using markov random fields," in *Proc. IEEE Int'l. Conf. Computer Vision*, pp. 1050–1057, Oct. 2009.
- [94] W. H. Deng, J. N. Hu, and J. Guo, "Extended src: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 9, pp. 1864–1870, 2012.
- [95] W. H. Deng, J. N. Hu, and J. Guo, "In defense of sparsity based face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.
- [96] A. Majumdar and R. Ward, "Classification via group sparsity promoting regularization," in *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing*, pp. 861–864, 2009.
- [97] J. J. Wang, J. C. Yang, K. Yu, F. J. Lv, T. Huang, and Y. H. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3360–3367, 2010.

- [98] Y. N. Liu, F. Wu, Z. H. Zhang, Y. T. Zhuang, and S. C. Yan, "Sparse representation using nonnegative curds and whey," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3578–3585, 2010.
- [99] Y. W. Chao, Y. R. Yeh, Y. W. Chen, Y. J. Lee, and Y.-C. Wang, "Locality-constrained group sparse representation for robust face recognition," in *Proc. IEEE Int'l Conf. Computer Vision Image Processing*, pp. 761–764, Sept 2011.
- [100] L. Zhang, M. Yang, and X. C. Feng, "Sparse representation or collaborative representation: Which helps face recognition?," in *Proc. IEEE Int'l. Conf. Computer Vision*, p. 471–478, Nov 2011.
- [101] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, no. 1, pp. 49–67, 2006.
- [102] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems*, 2002.
- [103] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in Neural Information Processing Systems*, 2009.
- [104] D. Lee and S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, no. 6775, pp. 788–791, 1999.
- [105] D. Lee and S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.
- [106] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [107] Q. Zhang and B. X. Li, "Discriminative k-svd for dictionary learning in face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2691–2698, June 2010.
- [108] Z. L. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1697–1704, 2011.
- [109] Z. L. Jiang, Z. Lin, and L. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, pp. 2651–2664, Nov 2013.

- [110] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.
- [111] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3501–3508, 2010.
- [112] M. Yang, D. Zhang, X. C. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Int'l Conf. Computer Vision*, pp. 543–550, 2011.
- [113] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *International Journal of Computer Vision*, vol. 109, no. 3, pp. 209–232, 2014.
- [114] C. F. Chen, C. P. Wei, and Y. C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2618–2625, 2012.
- [115] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2586–2593, 2012.
- [116] Y. Zhang, Z. L. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.
- [117] E. J. Candès, X. D. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, pp. 11:1–11:37, June 2011.
- [118] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, pp. 171–184, Jan 2013.
- [119] 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.
- [120] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji, "Discriminant sparse neighborhood preserving embedding for face recognition," *Pattern Recognition*, vol. 45, no. 8, pp. 2884 – 2893, 2012.
- [121] H. C. Zhang, Y. N. Zhang, and T. S. Huang, "Simultaneous discriminative projection and dictionary learning for sparse representation based classification," *Pattern Recognition*, vol. 46, no. 1, pp. 346 – 354, 2013.

- [122] Z. Z. Feng, M. Yang, L. Zhang, Y. Liu, and D. Zhang, "Joint discriminative dimensionality reduction and dictionary learning for face recognition," *Pattern Recognition*, vol. 46, no. 8, pp. 2134 – 2143, 2013.
- [123] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4 –37, Jan. 2000.
- [124] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
- [125] X. D. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, pp. 931–937, May 2009.
- [126] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 8, pp. 831–836, 1996.
- [127] X. D. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, pp. 383–394, Mar. 2008.
- [128] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition*, vol. 33, no. 11, pp. 1771 – 1782, 2000.
- [129] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 780–788, Jun 2002.
- [130] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. IEEE Int'l. Conf. Computer Vision*, vol. 2, pp. 1208 –1213 Vol. 2, oct. 2005.
- [131] X. D. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition," *Electronics Letters*, vol. 42, pp. 1089–1090(1), September 2006.
- [132] X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, pp. 328 –340, Mar. 2005.
- [133] S. C. Yan, D. Xu, B. Y. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, pp. 40–51, Jan 2007.
- [134] X. D. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, pp. 16–26, Mar. 2011.
- [135] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Advances in Neural Information Processing Systems*, 2006.

- [136] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, "Signal processing with compressive measurements," *IEEE J. Sel. Topics Signal Process.*, vol. 4, pp. 445–460, Apr. 2010.
- [137] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision Research*, vol. 37, no. 23, pp. 3311 – 3325, 1997.
- [138] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. Second IEEE Workshop Applications of Computer Vision*, p-p. 138–142, 1994.
- [139] A. Martinez and R. Benavente, "The ar face database," *CVC Technical Report*, June 1998.
- [140] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, pp. 1615–1618, Dec 2003.
- [141] X. T. Yuan and S. C. Yan, "Visual classification with multi-task joint sparse representation," in *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3493–3500, 2010.
- [142] N. Hurley and R. Scott, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, pp. 4723–4741, Oct 2009.
- [143] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, pp. 684–698, May 2005.
- [144] M. Yang, D. Zhang, D. Zhang, and W. S., "Relaxed collaborative representation for pattern classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2224–2231, June 2012.
- [145] L. El Shafey, C. McCool, R. Wallace, and S. Marcel, "A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 1788–1794, July 2013.
- [146] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1996.
- [147] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *UIUC Technical Report UILU-ENG-09-2215*, 2009.
- [148] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and Y. S. C., "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, pp. 1031–1044, June 2010.

- [149] A. Wagner, J. Wright, A. Ganesh, Z. H. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, pp. 372–386, Feb 2012.
- [150] X. D. Jiang and J. Lai, "Sparse and dense hybrid representation via dictionary decomposition for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, pp. 1067–1079, May 2015.
- [151] C. Geng and X. D. Jiang, "Face recognition based on the multi-scale local image structures," *Pattern Recognition*, vol. 44, no. 10C11, pp. 2565 – 2575, 2011.
- [152] C. Geng and X. D. Jiang, "Fully automatic face recognition framework based on local and global features," *Machine Vision and Applications*, vol. 24, no. 3, pp. 537–549, 2013.
- [153] Z. W. Miao and X. D. Jiang, "Interest point detection using rank order log filter," *Pattern Recognition*, vol. 46, no. 11, pp. 2890 – 2901, 2013.
- [154] B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: global versus component-based approach," in *Proc. IEEE Int'l Conf. Computer Vision*, vol. 2, pp. 688–694, 2001.
- [155] B. Heisele, P. Ho, J. Wu, and T. Poggio, "Face recognition: component-based versus global approaches," *Computer Vision and Image Understanding*, vol. 91, no. 1C2, pp. 6 – 21, 2003. Special Issue on Face Recognition.
- [156] B. Heisele, T. Serre, and T. Poggio, "A component-based framework for face detection and identification," *International Journal of Computer Vision*, vol. 74, no. 2, pp. 167–181, 2007.
- [157] J. Lai and X. Jiang, "Modular weighted global sparse representation for robust face recognition," *IEEE Signal Processing Letters*, vol. 19, pp. 571–574, Sept 2012.
- [158] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
- [159] E. Grave, G. Obozinski, and F. Bach, "Trace lasso a trace norm regularization for correlated designs," in *Advances in Neural Information Processing Systems*, 2012.
- [160] W. Deng, W. Yin, and Y. Zhang, "Group sparse optimization by alternating direction method," technical report, rice university, 2011.
- [161] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, pp. 237 – 260, 1998.

- [162] P. Zhao and B. Yu, “On model selection consistency of lasso,” *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, Dec. 2006.
- [163] R. He, W. S. Zheng, T. Tan, and Z. Sun, “Half-quadratic-based iterative minimization for robust sparse representation,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 36, pp. 261–275, Feb 2014.
- [164] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-pie,” *Image and Vision Computing*, vol. 28, no. 5, pp. 807 – 813, 2010.
- [165] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang, “Collaborative representation based classification for face recognition,” *arXiv: 1204.2358*, 2014.
- [166] Q. Shi, A. Eriksson, A. v. d. Hengel, and C. Shen, “Is face recognition really a compressive sensing problem?,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 553–560, June 2011.
- [167] R. Rigamonti, M. Brown, and V. Lepetit, “Are sparse representations really relevant for image classification?,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1545–1552, June 2011.
- [168] I. Tosić and P. Frossard, “Dictionary learning,” *IEEE Signal Processing Magazine*, vol. 28, pp. 27–38, March 2011.
- [169] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, “Supervised dictionary learning,” in *Advances in Neural Information Processing Systems*, 2008.
- [170] L. S. Z., “Face recognition based on nearest linear combinations,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 839 –844, jun 1998.
- [171] J. Ho, M. H. Yang, J. Lim, K. C. Lee, and D. Kriegman, “Clustering appearances of objects under varying illumination conditions,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. I–11–I–18 vol.1, June 2003.
- [172] Z. C. Lin, A. Ganesh, J. Wright, L. Q. Wu, M. M. Chen, and Y. Ma, “Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix,” in *In Intl. Workshop on Comp. Adv. in Multi-Sensor Adapt. Processing*, 2009.
- [173] J. Cai, E. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.
- [174] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The feret evaluation methodology for face-recognition algorithms,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 10, pp. 1090–1104, 2000.

- [175] A. Satpathy, X. D. Jiang, and H. Eng, "Human detection by quadratic classification on subspace of extended histogram of gradients," *IEEE Trans. Image Processing*, vol. 23, pp. 287–297, Jan 2014.
- [176] J. Ren, X. D. Jiang, and J. Yuan, "Noise-resistant local binary pattern with an embedded error-correction mechanism," *IEEE Trans. Image Processing*, vol. 22, pp. 4049–4060, Oct 2013.
- [177] A. Satpathy, X. D. Jiang, and H. Eng, "Lbp-based edge-texture features for object recognition," *IEEE Trans. Image Processing*, vol. 23, pp. 1953–1964, May 2014.
- [178] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, pp. 2233–2246, Nov 2012.
- [179] M. Yang, L. Zhang, and D. Zhang, "Efficient misalignment-robust representation for real-time face recognition," in *Proc. European Conf. Computer Vision*, vol. 7572, pp. 850–863, 2012.