The Hong Kong Polytechnic University

Department of Computing

Robust and Efficient Face Recognition via Adaptive Masking and Dictionary Learning

by

Zhizhao Feng

A thesis submitted in partial fulfillment of the requirements

for the Degree of Master of Philosophy

February 2012

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree of diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Zhizhao Feng (Name of Student)

Abstract

Biometrics technologies have been more and more widely used in our daily life to identify a person by investigating his/her physiological and behavioral characteristics. Among various biometrics identifiers, face as a distinctive and easy to use biometrics identifier has been widely studied for more than thirty years. However, there are still many challenging problems remaining in face recognition. In this thesis, we propose two coding based robust and efficient face recognition schemes which aim to solve the face occlusion problem and discriminative feature extraction problem, respectively.

Face occlusion refers to that the query images are occluded partially by glasses, scarves, or irrelevant images. The occlusion sometimes covers a large part of the frontal face image, which greatly compromises facial feature extraction in conventional methods, leading to failure in face recognition and limiting their applications to practical systems. It is known that occluded pixels usually have high coding errors when representing a face image over the non-occluded training samples. Based on this fact, we propose a novel coding residual map learning scheme for fast and robust face recognition, namely *Fast Robust Face Recognition via Coding Residual Map Learning based Adaptive Masking* (CRMLAM). A dictionary is learnt to code the training samples, and the distribution of coding residuals can be learnt simultaneously. A residual map can then be obtained to detect the occlusions by adaptive thresholding. Finally the face image can be identified robustly by masking the detected occlusion pixels from face representation. The occluded pixels can be approximately located through this method, and thus the recognition rate can be greatly increased comparing with some state-of-the-art face recognition methods.

In face recognition, the computational cost is always an important factor to be considered. Face image usually lies in a high dimension space. Although some prevailing face recognition methods can achieve competing recognition result, the large amount of

computational cost greatly reduces their availability. In our proposed method, both the face coding residual and the face coding coefficients are modeled by l_2-norm, and thus the time consumed in face representation and occluded pixel detection is low. By our experiments on benchmark and large scale face databases, the total amount of time cost for recognizing one query image is normally less than one second under the Matlab programming environment, which is very fast compared to state-of-the-art robust face recognition methods. Meanwhile, the face recognition accuracy by our method is very competitive.

The high dimension of face image not only leads to high computational cost, but also prevents the discriminative features from being used for face recognition. In fact, there is much trivial information in the face image which is not desirable in face recognition. Many dimensionality reduction methods have been proposed to solve this problem, while the dictionary learning methods can also be used to reduce the redundant information for a more accurate face representation. Dimensionality reduction and dictionary learning are often considered as two separate steps; in this thesis, we propose a joint learning scheme of dimensionality reduction and dictionary learning, namely *Joint Discriminative Dimensionality Reduction and Dictionary Learning* (JDDRDL). A face projection matrix and a face representation dictionary are learnt simultaneously by one objective function. By JDDRDL, it is expected that the face features could lie in a more discriminative low dimensional space, where a more representative dictionary can be used to code the face features. Since discriminative information is enhanced in both projection and dictionary learning, the proposed method can better handle the small sample size problem in face recognition. When the number of training sample is insufficient, the recognition rate of many dimensionality-reduction or dictionary-learning based face recognition methods will drop a lot. In comparison, the proposed JDDRDL is still able to achieve satisfying recognition result by exploiting effectively the training information.

The major contributions of this thesis are summarized as follows:

(1) An efficient and robust face recognition scheme is proposed by learning a dictionary and a coding residual map from the training samples, and coding the query sample over the learnt dictionary with adaptive masking. The proposed method is robust to face occlusion but with a low computational cost;

(2) A joint discriminative dimensionality reduction and dictionary learning scheme is developed, which is more robust to the small sample size problem and achieves better face recognition results than state-of-the-art methods.

Publications

The following papers, published or submitted, are the partial outputs of my MPhil studies in PolyU.

1. Lei Zhang, Meng Yang, **Zhizhao Feng**, David Zhang, "On the Dimensionality Reduction for Sparse Representation based Face Recognition", ICPR 2010.

2. **Zhizhao Feng**, Mang Yang, Lei Zhang, David Zhang, "A Fast Robust Face Recognition Approach via Coding Residual Map Learning based Adaptive Masking", submitted to ECCV 2012.

3. **Zhizhao Feng**, Meng Yang, Lei Zhang, Yan Liu and David Zhang, "Joint Discriminative Dimensionality Reduction and Dictionary Learning for Face Recognition", submitted to Pattern Recognition.

Acknowledgements

This two years' MPhil study is now going to the end, I'm very glad to take this opportunity to express my appreciation to all the people concerning and supporting me in the past two years.

First of all, I'd like to say thank you to my chief supervisor, Dr. Lei Zhang, for his patient and professional supervision on my research work. In the past two years, Dr. Zhang had shown me the scientific idea and unexhausted passion that a researcher should own as well as the endeavor which a researcher should devote to exploiting the unknown field. Dr. Zhang's diligence and meticulosity really impact me a lot. I will always keep them in my mind in my future work and life.

Besides Dr. Zhang, I'd also like to express my gratitude to other professors in my department, they are Prof. David Zhang, Dr. Yan Liu, Dr. Eric Lo and Dr. Ken Yiu. They all gave me helpful advice during my study. My colleagues also gave me great support and kind help in the past two years, they are Dr. Lin Zhang, Dr. Bo Peng, Dr. Maggie Guo, Dr. Denis Guo, Mr. Meng Yang, Mr. Kaihua Zhang, Mr. Pengfei Zhu, Mr. Jin Xie, Mr. Andy Wang, Mr. Xiaofeng Qu, Mr. Jinghua Wang and Miss. Feng Liu. I'd also like to appreciate my friends, without their encouragement and help, I would never finish this journey. They are Ken Luo, Allen Hu, Allen Gu, Libra Huang, Roy Ye, Maggie Zhu, Owen Xiao, Yanling Pan, Andy He, Shirley Sun, Duncan Yung, Petrie Wong, Victor Liang, Yu Li, Jianguo Wang, Yufei Wang, Gene Wu, Dr. Edward Teng and Dr. Jackie Wu.

Lastly, my deepest gratitude goes to my parents and family for their understanding and great mental support, they will not read this thesis, but their support is most critical for me to finish this journey.

Table of Content

List of Figures

List of Tables

Chapter 1. Introduction

1.1 A Brief Introduction to Faces as Biometrics Identifier

Personal identification authorization is now widely used in many aspects of daily life, such as access control, passenger clearance and crime investigation among other things. The most common personal identification methods are ID cards, passports, keys, passwords and other materials or property which can identify an owner or a certain authorized person. However, these methods have two main drawbacks. First, they can be either lost, so that even if one is the supposed authorized person, access will still be denied, or forgotten. Statistics shows that on average a person usually uses about 30 passwords in his/her lifetime. With so many passwords one may get confused easily. If a password is lost or forgotten, inconvenience and annoyance will be resulted. Second, these documents can be duplicated and passwords may be stolen. With duplicated documents or stolen passwords, unauthorized people can intrude on others' privacy or gain unauthorized access.

Biometric based authorization methods [1], however, can overcome above drawbacks in a reliable way. Biometrics methods use unique inherent physical or behavioral characteristics of human beings to determine personal identity [1, 2, 3]. It is known that people have some traits that are sufficiently different for individuals in the relevant population such that they can be distinguished from one another. The human face, as a natural choice of a biometric element for identification, is both easily accessible and socially acceptable [4]. In addition, there are advances in the computational powers of modern computers which allow application of more complex algorithms. As a popular biometrics identifier, faces are universal, and owned by all human beings. Also, faces can be easily collected from daily life, which provides great convenience for algorithm testing

and system building. Unlike fingerprints, which are usually linked with crime investigation, face recognition is widely accepted by users.

1.2 Procedure of a Face Recognition System

As an effective and reliable identification method, face recognition remains a hot topic in the recent decades [5-17]. A typical face recognition system includes the following parts [51]:

(1) Face detection. It aims to locate the face region from an image scene, or a sequence of images. It is the fundamental part of a recognition system, since further processing steps are based on region detection. The region located should feature mostly a front face part, with less background, so that the computation cost can be reduced considerable if detection is accurate.

(2) Preprocessing. It aims to reduce noises and align or normalize face images, in order to facilitate feature extraction.

(3) Feature extraction. Face images usually lies in a high dimensional space, which induces high computation cost and prevents robust and fast face recognition. Feature extraction aims to extract important face features from a high dimensional face image. Those features can well represent essential information on face images, as well as remove trivial information which may compromise recognition.

(4) Feature matching. After features from the original face image are extracted, they will be used in matching different classes of face images in the dataset. If certain criteria are met, the face can be classified into some class.

Fig. 1.1 shows the general procedure of a face recognition system.

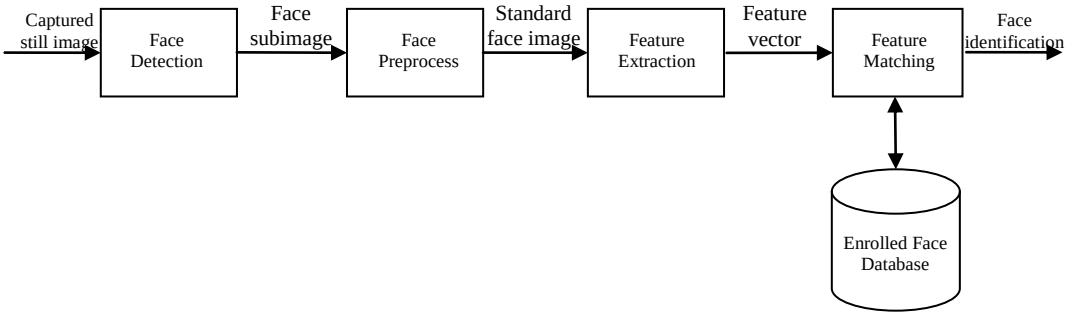


Figure 1.1: A general work flow of a face recognition system.

In practice, there are usually two scenarios for face comparison, which are face recognition/identification and face verification/authorization [51, 52, 53].

(1) Identification. Face identification means that there are several classes of face image with known identities in a face database. Our task is to determine the identity of a new test image, which belongs to a certain class in the gallery set but is not the exactly same face image shown in the Fig. 1.2. Face identification can be difficult without any previous knowledge of the test sample, but its class has to be found out in a huge gallery set. Some criteria schemes need to be developed and matching the probe image and each image from the gallery set requires fast and robust algorithms.

(2) Verification. Like identification, there is a set of registered face images with known identities or classes in face verification. However, face verification is sometimes easier since the probe image has to "claim" its class. After the "claim", the registered images from the same class in the gallery set will be separately taken out to match the probe image. If their difference is smaller than certain criteria, the probe image is assumed to be "genuine"; otherwise, it is assumed to be an "imposer". The Fig. 1.3 shows a verification system. Since verification is only needed in comparison, it is much more efficient than identification. However, criteria determination remains a very controversial research topic.

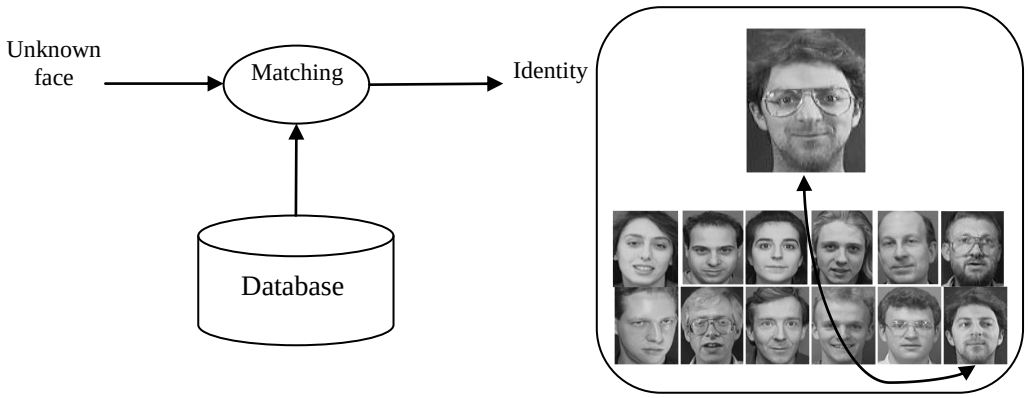


Figure 1.2: Demonstration of a face identification system.

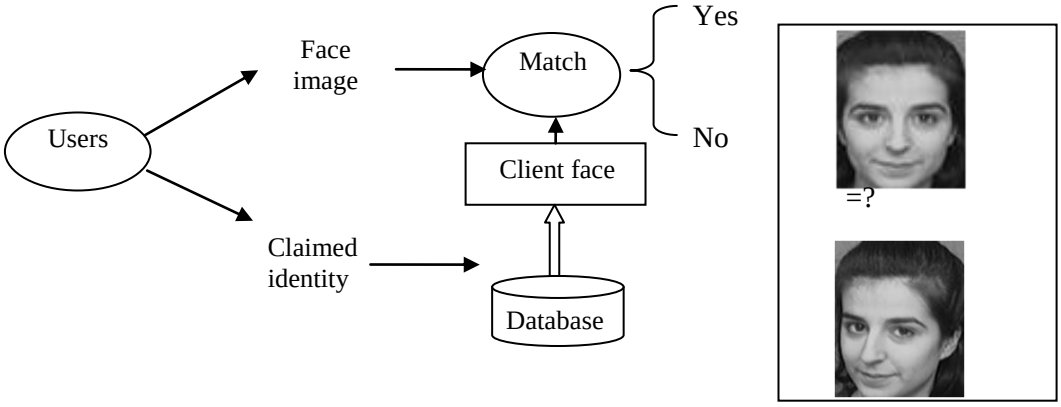


Figure 1.3: Demonstration of a face verification system.

This thesis aims to illustrate an effective and robust face recognition scheme with its focus on face identification instead of face verification.

1.3 Challenges of Face Recognition

Although face recognition has been studied for more than thirty years, there are still many problems unsolved. The most challenging ones are summarized below:

(1) High dimensionality. As discussed above, face images usually lie in a high dimensional space. High dimensionality demands a lot of computational effort, which makes on-line processing impossible. The high dimensionality also generates noises and contains a lot of irrelevant and redundant information, which may compromise recognition.

(2) Face occlusion. It always presents a challenging problem to face recognition. However, faces can be occluded by glasses, masks or scarves which cover the geometric features and leave only certain parts visible, thus causing great difficulty in feature extraction.

(3) Illumination/pose/expression variation. When a face is exposed in different light conditions, face images can vary considerably. Illumination can come from different directions, under different sources, which makes face images also appear differently. A person can display different face images from different angles. Expression variation can also cause difficulty in face recognition. For instance, smiling and being scared make quite different face expressions.

(4) Small sample size. Traditional face recognition methods usually require a training dataset to contain supposedly enough samples so that they can extract features efficiently. However, due to storage problem, samples of each subject sometimes may be insufficient to enable the satisfying performance of those algorithms.

1.4 Major Approaches to Face Recognition

Face recognition has been a hot topic in the past 30 years and due to its wide array of practical applications, there are many approaches to face recognition. Zhao *et al.* [54] summarized the major approaches into three categories, namely holistic feature based methods, local feature based methods and hybrid methods.

1.4.1 Holistic Feature Based Methods

Holistic feature methods are also known as appearance-based methods, in which a face image is extended into a high dimension vector. As the face image is represented by a vector, some feature extraction methods can be applied to the face vector to extract some low-dimensional features, and then certain classification metrics can be used to determine the identity of the face image. Principle Component Analysis (PCA) [55] is one of the most

representative and popular methods in the holistic class, in which a face image is represented as a linearly weighted sum of a set of orthogonal basis vector. A view-based and modular eigenspace method is thus proposed to handle the larger dataset problem [18]. [19] also proposed a nonlinear PCA auto-encoder multilayer neutral network using a nonlinear projection to improve PCA in terms of better reconstruction. Since nonlinear PCA requires nonlinear optimization and sometimes it may be trapped in over-fitting, kernel based PCA is proposed to overcome this problem [20, 21, 22, 23]. The Gabor-based Kernel Principal Component Analysis (GKPCA) method [23] integrates the Gabor wavelet representation of face images and the kernel PCA method to solve the pose and illumination variation.

Linear Discriminative Analysis (LDA) [24] is also a representative method in the holistic class. When there are multiple samples in each class, LDA can be used to expressly provide discrimination among classes. Aiming to deal with small sample size problem, Zhao *et al*. [25, 26] introduced a subspace linear discriminative analysis method, in which PCA is used to extract the face subspace, and then LDA is used to obtain a linear classifier, which is able to solve the problem of over-fitting effectively. But [27, 28] argued that PCA projection may cause to discard null space which may contain significant discriminatory information. To prevent this from happening, direct LDA (D-LDA) methods without a separate PCA step have been presented [27, 28]. In such a framework, training samples are processed in the original high dimensional space to avoid any loss of significant discriminatory information. Zuo *et al*. [29] also introduced a fast feature extraction technique before applying LDA, namely bidirectional PCA (BDPCA) plus LDA (BDPCA + LDA), which performs LDA in the BDPCA subspace to avoid any loss of discriminative information in dimension reduction. Like PCA, some kernel based LDA methods [30, 32] are also proposed to improve the LDA scheme. Lu *et al*. [30] proposed to use the polynomial function, radial basis function (RBF) and multilayer perceptrons to form a kernel function of LDA. However, the performance of the kernel method is highly

dependent on the kernel and its parameters. P. C. Yuen *et al*. [31] proposed to learn a kernel matrix via maximizing the difference between inter-class and intra-class similarities and introduced an "ILLUM" kernel model using the scattered data interpolation technique, which can greatly improve the performance.

Some recently developed representation based methods, such as SRC [15], CRC [35], receive robust and efficient results. Since our method is also based on the representation based holistic method, the above representative holistic schemes will be discussed in the next chapter.

1.4.2 Local Feature Based Methods

Feature based methods focus on the geometric features of a face structure, such as eyes, noses, chins and mouths. The features are extracted from face images and compared with templates and the identity can be determined. Compared with holistic feature based methods, local feature based methods are not sensitive to illumination variation and the expression change; since they mainly rely on the geometric features. However, the performance of feature based methods is highly related to the geometric feature extraction algorithm. If feature extraction fails to locate the needed features, local feature based methods will also fail to identify samples. Local Feature based methods are more sensitive to pose change. In the early years, some local feature based methods are like template matching. Bruneli *et al.* [36] automatically selected a set of 4 feature templates (i.e. eyes, noses, mouths and the whole face) in each training face image. For each query image, its regions are compared with the same regions through normalized cross correlation. The decision is made via the sum of matching score of each region. However, this method relies on templates heavily. A more efficient feature based method was proposed by Cox *et al*. [37], in which the distance function is constructed based on local second order statistics estimated by modeling training samples as a mixture of normal densities. To reduce the complication of computation in [37], Hjelmas [38] proposed a face recognition scheme

based on local feature extraction. The interesting feature points are located by Gabor filters, and most of them contain high information content and a feature vector consisting of Gabor coefficients will be extracted from each of those points. [39] applied a block-based discrete-cosine transform (DCT) to extract local features of face images, hoping that the compactness of representation is optimal. The DCT coefficients of each block are concatenated to construct feature vectors. The Local Binary Pattern (LBP) is also used to describe the face images in face recognition due to its invariance in monotonic gray-level changes and computational efficiency. Timo Ahonen *et al*. [40] proposed a face recognition scheme by using the LBP feature to describe face images. The face images are firstly separated and divided into several areas where the LBP feature distributions are calculated and then concatenated into a feature vector which can be used to describe faces.

1.4.3 Hybrid Methods

Hybrid methods use both holistic based and feature based methods. They usually use face shapes or its gray-level information to construct a face model, and PCA and LFA are applied to the model to extract features. It is argued that hybrid methods have the advantages of both holistic feature based methods and local feature based methods. However, a face model in hybrid methods is critical to its performance. If a face model is not suitable, identification results will be highly affected. A typical hybrid method is proposed in [41]. Lanitis *et al*. build two models to represent face outlines using 152 points and 160 training samples from 20 subjects. One is the shape model from geometric features, and the other is the grey-level model learnt from PCA. It is believed that the two models can represent most face outlines. For each query image, its grey model parameters and shape parameters are extracted and fully reconstructed from those models to determine its identity. However, [41] only achieves good results when there is not much variation in the database. Otherwise, the model cannot represent all samples accurately. Another popular hybrid method was proposed by R. Huang *et al*. [42] using Markov Random Fields (MRF). In Huang's method, face images are divided into several small patches and for

each patch and its ID, a MRF model is learnt and used to represent the relationship between them. Two compatibility functions are also learnt from training data, as well as the MAP solution to them. Query images are also divided into several patches as training samples, and the identity is determined through voting from those patches. Since compatibility functions and solutions can be obtained offline, recognition speed is quite fast by using Huang's method. Lawrence *et al*. [43] also proposed a hybrid neural network method, using a self-organizing map neutral network and a convolution neural network. A similar work was proposed by Gorodnichy *et al*. in [44] and Kang *et al*. in [45]. In [44] they used adaptive logic networks to extract facial features and a hybrid of hidden Markov chain and neural networks was applied to [45].

1.5 Contribution and Outline of this Thesis

From the above introduction, it can be seen that face recognition is of great value to practical application. A robust face recognition system can give us great convenience. Recently, researchers have developed many face recognition algorithms and systems. However, there are still unresolved problems. Based on the recent development of sparse representation classification (SRC) [15], this thesis exploits a face recognition scheme which aims to deal with occlusion problem in a rapid speed and lower computation cost. An error map learnt from the original unmasked face image is used to determine occlusion points in the test process and the located occlusion points are discarded in reconstruction and recognition steps. The new l_2 norm constrain is also adopted to design the optimization function, which turns the occlusion problem into a least square problem, thus inducing much less computation cost than the l_1 SRC scheme. Another important contribution of this thesis is that a robust joint discriminative dimension reduction and dictionary learning scheme (JDDRDL) is proposed. Dimension reduction is an important preprocessing step in face recognition, such as PCA and LDA. However, as shown in the following sections, all traditional methods have drawbacks and cannot fit the SRC model. Dictionary learning is a

popular research topic of face recognition and attracts more and more attention. In this thesis, dimensional reduction and dictionary learning are combined in an attempt to obtain an appropriate dimensional reduction matrix and a representative dictionary and receive satisfactory recognition results. This JDDRDL method can also handle face occlusion. The proposed methods both depend on the representation based face recognition methods. It is true that the representation based methods require that the testing images should be able to be linearly combined by the training samples, otherwise those methods will fail (e.g. pose variations, misalignment). As a result, the proposed methods will also be verified in suitable experiments. The detailed algorithms and experiment results will be discussed in the following sections. The main contributions of this thesis are as follows:

(1) Propose a simple but efficient outlier pixels detection scheme to detect outlier pixels (e.g., occlusion, disguise) in face image.

(2) Based on the outlier detection algorithm and the sparse representation model, an efficient face recognition algorithm, namely CRMLAM, is developed to handle face occlusion at a low computation cost. The running time for one query image is less than one second but the recognition rate is quite satisfactory.

(3) Propose to learn the discriminative dimension reduction matrix and dictionary jointly to represent face images in a low dimensional space. This joint learning scheme, namely JDDRDL, fills the gap between dimension reduction and dictionary learning by combining the two techniques to better fit the sparse representation model.

(4) Apply JDDRDL to small sample size face recognition and face occlusion, verifying that the discriminative dimension reduction matrix and dictionary could well overcome the above challenges of face recognition.

The rest of this thesis is presented as follows:

Chapter 2 briefly reviews the algorithms used or referred to in this thesis, including PCA, LDA, SRC and CRC.

Chapter 3 presents a fast and robust face recognition scheme to handle face occlusion. Motivation and details, construction of error map and selection of parameters are discussed Experiment results are illustrated to compare prevailing methods.

Chapter 4 discusses the importance of introducing the dimensional reduction matrix and dictionary and proposes the joint discriminative dictionary learning and dimensional reduction scheme. The background of dictionary learning and dimensionality reduction is introduced and experiments are conducted on benchmark databases to illustrate the effectiveness of the scheme.

Our findings are summarized and future work is proposed in Chapter 5.

Chapter 2. Related Works on Representation Based Face Recognition

As discussed in Chapter 1, holistic feature based methods are among the most important face recognition techniques. They extend a face image to a vector and do not consider the face's geometric information, turning recognition into vector/matrix analysis, so that many mathematic analysis tools can be directly applied to analysis and recognition procedures. PCA [55] and LDA [24] are two representatives of holistic methods. Both methods aim to find some projection basis vectors for dimensional reduction and recognition is applied into low dimensional feature space. In this thesis, SRC and CRC are classed as the presentation based holistic methods since PCA and LDA use some kinds of basis vectors to project or reconstruct face images. SRC and CRC use face images as the templates to reconstruct a test image, and recognition is performed on reconstruction error. And based on this, many related methods have since been proposed, such as the l_1-graph for image classification [46], kernel based SRC [47, 48], robust sparse coding [49], robust alignment with spares and low rank decomposition [63]. In the following section, PCA, LDA, SRC and CRC are discussed in details.

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is also known as eigenface, which extracts eigenvectors of high dimensional data to model linear variance, and projects high dimensional data into a linear subspace spanned by a low dimensional feature through leading eigenvectors of the data's covariance matrix, aiming to seek a subspace with maximized variance. Suppose there is a set of training samples $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2\ldots, \boldsymbol{x}_k]$ with dimension of m, the covariance of the zero mean samples $\boldsymbol{X}$ is known as $\boldsymbol{C}=\frac{1}{\boldsymbol{k}}*(\boldsymbol{X}-\bar{\boldsymbol{X}})*(\boldsymbol{X}-\bar{\boldsymbol{X}})^T$, where $\bar{\boldsymbol{X}}$ is the global mean value of $\boldsymbol{X}$. According to the

matrix theory, a set of eigenvectors and the corresponding eigenvalues can be found, so that $\boldsymbol{C\Phi}=\boldsymbol{\Phi\Lambda}$, where $\boldsymbol{\Phi}=[\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_{2,\ldots,}\boldsymbol{\Phi}_n]$ and $\boldsymbol{\Lambda}=[\alpha_1,\alpha_2,\ldots,\alpha_n]^T$ are eigenvectors and corresponding eigenvalues respectively. The eigenvectors corresponding to the top $k<n$ eigenvalues span a low dimensional feature space. For any vector $\boldsymbol{y}$, it can be projected into a low dimensional space, so that the vector is represented by a linear combination of the eigenvectors with weights.

2.1.1 Calculate Eigenfaces

Usually, a face image is of size $M*N$, and it can be transferred through row concatenation into a long vector of $(M*N)*1$. As a result, face image is deemed as a point in a high dimensional space. All face images in a training dataset form a face image space. As stated above, the aim of PCA is to seek a set of orthonormal basis vectors which best characterize the distribution of face images in a face images space, for a certain projection basis vector $\boldsymbol{u}_k$, it should satisfy the following objective function

$$\lambda_k = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{u}_k^T \hat{\boldsymbol{X}}_n)^2 \tag{2-1}$$

where $\hat{\boldsymbol{X}}_n = \boldsymbol{X}_n - \boldsymbol{E}(\boldsymbol{X}_n)$, the original face image is subtracted by the global mean.

Since it is required that the basis vector should be orthornormal, the constraint is imposed as

$$\boldsymbol{u}_i^T \boldsymbol{u}_k = \begin{cases} 1, & \text{if } i = k \\ 0, & \text{otherwise} \end{cases} \tag{2-2}$$

It is found that $\boldsymbol{u}_k$ is the eigenvectors of C and λ_k is the corresponding eigenvalue. The theory behind PCA is the eigenvectors and eigenvalues of the matrix. There are also many existing fast algorithms to solve the problem of eigenvector and eigenvalue.

2.1.2 Face Recognition based on Eigenfaces

PCA is a simple and efficient dimension reduction method. The leading eigenvectors extracted from original samples can reconstruct an original image efficiently. Face images can be represented by linear combination by the eigenvectors. An intuitive way to apply PCA to face recognition is to compare the weight of the test image with that of the known training data. In [55], a face recognition scheme was proposed based on eigenfaces. The scheme is reviewed below:

(1) A set of training data, each class of which contains several samples, is imported; then eigenvalues and eigenvectors are calculated from the training dataset;

(2) The leading *k* eigenvalues and the corresponding eigenvectors are chosen. In practice, the value of *k* does not need to be large since perfect reconstruction is not necessary in recognition. A large *k* would greatly increase computational complexity.

(3) When eigenvalues and eigenvectors are chosen, the training dataset is projected into the subspace spanned by eigenvectors and the weight of each eigenvector is calculated. The corresponding distribution of each known sample has thus been obtained from the training dataset.

(4) A probe image is projected into the subspace formed by the aforementioned eigenvectors so that the weight coefficients are obtained and the probe image is linearly reconstructed by eigenvectors. Recognition is performed by comparing the weight coefficients between the probe image and each sample in the training dataset. The identity is determined by the class which leads to minimum difference.

2.1.3 Advantage and Limitation of PCA

PCA is a simple scheme for face recognition. It only considers the leading eigenvectors of the training samples which span a subspace where the probe image can be compared with the training data to determine the identity. Computation cost is low and implementation is easy and therefore PCA attracts much attention from researchers. PCA also performs satisfactorily when dealing with cases of simple face recognition. However, PCA only preserves the global structure but in face recognition, sometimes local structure is more

important. As a result, PCA suffers whenever there is illumination change or expression variation. Although PCA is no longer a prevailing direct face recognition method, it is still widely used in dimension reduction because noises and trivial information can be removed from face images after projection, which facilitates feature extraction and reduces the computation complexity.

2.2 Linear Discriminative Analysis (LDA)

LDA [24], also known as Fisherface, is another popular dimensional reduction method widely used for face recognition and other pattern recognition issue. Unlike PCA, LDA adopts the Fisher Criteria to separate different features. In the following section, the algorithm and theory of LDA are introduced.

2.2.1 The Fisher Criteria of LDA

As mentioned above, PCA does not perform well when dealing with illumination change since variation causes features to vary considerably. In fact, illumination change is a common problem in face recognition. A typical example can be seen in Fig. 2.1, in which the cases of illumination change come from the Extended Yale-B database [56]. The four pictures feature the same person. However they look significantly different. The first picture (starting from left) shows the light directs towards the person; in the second picture, the dominating light comes from the top. In the third and fourth pictures, the light comes from right and left, respectively.

Figure 2.1: The same person under different light sources

LDA aims to learn a series of vectors which can project face images into a low dimensional space insensitive to illumination and expression change. Unlike PCA, LDA is a supervised dimensional reduction method which utilizes class information from the training data, and focuses on minimizing the intra distance between samples in the same classes while maximizing the inter distance between different classes. This is called the Fisher Criteria. The Fisher Criteria increases the inter scatter and decreases the intra scatter. In the format of mathematic function, the LDA objective function can be represented as follows:

$$\boldsymbol{W} = \arg\max_{\boldsymbol{W}} \frac{\boldsymbol{W}^T \boldsymbol{S}_B \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_W \boldsymbol{W}} \tag{2-3}$$

where

$$\boldsymbol{S}_B = \sum_{i=1}^{k} N_i (\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}})(\boldsymbol{x}^{(i)} - \bar{\boldsymbol{x}})^T \tag{2-4}$$

$$\boldsymbol{S}_W = \sum_{i=1}^{k} \sum_{j=1}^{N_i} (\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}})(\boldsymbol{x}_j^{(i)} - \bar{\boldsymbol{x}})^T \tag{2-5}$$

$\boldsymbol{W}$ is the projection vectors we want, $\bar{\boldsymbol{x}}$ is the global sample mean vector, $\boldsymbol{x}^{(i)}$ is the mean vector of the *i*th class samples, $\boldsymbol{x}_j^{(i)}$ is the *j*th sample in the *i*th class, N_i is the number of samples in the *i*th class and *k* is the total number of classes. $\boldsymbol{S}_B$ is called the between-class scatter matrix and $\boldsymbol{S}_w$ is called the within-class scatter matrix. The objective function aims to minimize $\boldsymbol{S}_w$ while maximizing $\boldsymbol{S}_B$, which is consistent with the Fisher Criteria. To achieve this goal, LDA uses the class label of the training set to find vectors which span low dimensional features. Since vectors have the preliminary class information, the projection is able to discount illumination variation and expression change. However, if class information is unclear in the training set, LDA will not be applicable.

2.2.2 Calculate Fisherface

It is assumed that there are all together *k* classes in the training set and LDA learns *k*-1 vectors to project and separate high dimensional face images. First, the between-class

scatter matrix $\boldsymbol{S}_B$ and the within-class scatter matrix $\boldsymbol{S}_w$ are calculated by Eq. (2-4) and Eq. (2-5). After projection, the between-class scatter is increased and the within-class scatter is decreased. An intuitive method to achieve this is to optimize the following objective function:

$$\boldsymbol{W}_{opt} = \arg\max_{w} \frac{\left\| \boldsymbol{W}^T \boldsymbol{S}_B \boldsymbol{W} \right\|_2^2}{\left\| \boldsymbol{W}^T \boldsymbol{S}_W \boldsymbol{W} \right\|_2^2} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n] \tag{2-6}$$

To maximize the objection function, the numerator should be as large as possible while the denominator should be as small as possible, which coincides with our objective. The solution to the function, $\boldsymbol{W}_{opt}$=[$\boldsymbol{w}_1$, $\boldsymbol{w}_2$,…, $\boldsymbol{w}_n$] should be the set of generalized eigenvectors of $\boldsymbol{S}_w$ and $\boldsymbol{S}_B$ associated with the n largest generalized eigenvalues {λ_i, i=1,2,…n}, i.e.

$$\boldsymbol{S}_B \boldsymbol{W}_i = \lambda_i \boldsymbol{S}_W \boldsymbol{W}_i, \ \ i = 1, 2, .., n \tag{2-7}$$

According to the matrix theory, there are k-1 nonzero eigenvalues at most, so n cannot be larger than k-1.

2.2.3 Advantage and Limitation of LDA

As mentioned in Section 2.2.1, one major advantage of LDA is that the Fisher Criteria can handle illumination and expression variation problem satisfactorily. The discriminative criteria ensures that the intra class distance is small while the inter class distance is large after projection into a low dimensional subspace. When dealing with the linear separatable cases, LDA usually achieves better result than PCA according to a large amount of experiment results [39].

However, when calculating the projection vectors W_i, the number of associated eigenvalues and eigenvectors is constrained by the number of classes in the training dataset. In other words, the reduced dimension size cannot exceed the number of classes in the training dataset, which is an apparent limitation of LDA. Another limitation of LDA is that

LDA assumes all features or patterns are linearly separable, but when the real case infringes this assumption, its performance will degenerate dramatically.

2.3 The Sparse Representation Based Classification (SRC)

The recently proposed SRC [15, 34] scheme quickly attracts researchers' attention as a robust holistic face recognition method. SRC deems each training face image as one column of dictionary matrix of representative samples, and an input testing image is represented as a sparse linear combination of these sample images via a certain constraint. Based on this model, other improved methods have been proposed [57-61].

2.3.1 Brief Introduction to the SRC Model

Sparse representation refers to the use of a relatively small number of bases/atoms of a dictionary to represent a certain signal. In the signal processing community, sparse representation was first used to reconstruct the signal on an over-complete dictionary. However, researchers did not realize its discriminative nature. In [15], it is exploited for its powerful discriminative ability in face recognition. The test sample is represented in an over-complete dictionary whose base elements are the original training samples. It is supposed that there are sufficient samples in each class. It is possible that the test sample can be linearly represented on the subset of a dictionary which only contains training samples from the class where the test sample belongs. Intuitively, this representation is naturally sparse since it is mainly reconstructed by a small portion of the dictionary (Fig. 2.2). It is also argued that the sparsely represented test sample can be recovered efficiently via l_1-minimization, which provides mathematic formalization to describe this model, which will be illustrated in the following section. The test image can be identified through its sparse representation coefficients. Ideally, the class where the test image belongs contains all nonzero entries. However, due to noises and modeling errors, there will always be small nonzero entries associated with other classes. To determine the identity, SRC

adopts global representation but investigates how well each class can reconstruct the test sample through its associated coefficients. The class which leads to the smallest reconstruction error is believed to be the right class where the test sample belongs. To better describe the SRC model, mathematic expression is used.

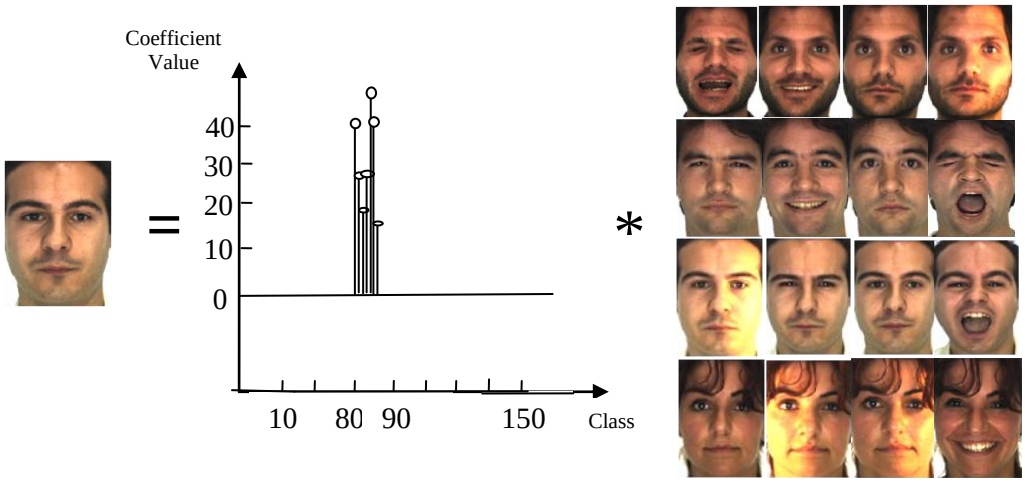


Figure 2.2: SRC represents a test image (left one) with an ideal sparse linear combination of all the training images (right ones).

Let $\boldsymbol{A}_i=[\boldsymbol{s}_{i,1}, \boldsymbol{s}_{i,2},\ldots,\boldsymbol{s}_{i,n}]\in R^{m*n}$ denote the training samples of the ith class, where $\boldsymbol{s}_{i,j}$ j =1,2,…,n, is an m-dimension vector stretched by the ith class. For each test sample $\boldsymbol{y}_0\in R^m$ from the ith class, this test sample can be well approximated through the linear combination within the training samples from $\boldsymbol{A}_i$, that is a coefficient vector $\boldsymbol{\alpha}_i=[\boldsymbol{\alpha}_{i,1},\boldsymbol{\alpha}_{i,2},\ldots,\boldsymbol{\alpha}_{i,n}]^T\in R^n$ can be found so

$$\boldsymbol{y}_0 = \sum_{j=1}^{n_i} \boldsymbol{\alpha}_{i,j}\boldsymbol{s}_{i,j} = \boldsymbol{A}_i\boldsymbol{\alpha}_i \tag{2-8}$$

Suppose there are k classes, and $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_k]$ is the concatenation of the n training samples from all the k classes, where $n=n_1+n_2+\ldots+n_k$ denotes the total number of classes of the samples. If $\boldsymbol{A}$ is used to represent the input test sample $\boldsymbol{y}_0$, then $\boldsymbol{y}_0=\boldsymbol{A}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}=[\alpha_1, \alpha_2,\ldots,\alpha_n]^T\in R^n$ is the corresponding coefficients. Since $\boldsymbol{y}_0$ comes from the ith class, and $\boldsymbol{y}_0=\boldsymbol{A}_i\boldsymbol{\alpha}_i$, that all other elements except $\boldsymbol{\alpha}_i$ are zero will be an intuitive solution to $\boldsymbol{\alpha}$, while $\boldsymbol{\alpha}_i$

holds significant values. In other words, the sparse non-zero elements in $\boldsymbol{\alpha}$ can encode the identity of the test sample $\boldsymbol{y}_0$, which is the essential idea of sparse representation based classification in [15]. In other word, the coefficient is coded via

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha}\{\|\boldsymbol{y}_0 - \boldsymbol{A}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1\} \tag{2-9}$$

where λ is a regularization parameter.

Recognition is based on the least reconstruction error from each class.

$$\text{identity}(\boldsymbol{y}_0) = \arg\min_{i}\|\boldsymbol{y}_0 - \boldsymbol{A}_i\hat{\boldsymbol{\alpha}}_i\|_2^2 \tag{2-10}$$

where $\hat{\boldsymbol{\alpha}}_i$ is the sub-coefficients associated with the ith class in $\hat{\boldsymbol{\alpha}}$.

Since the system of equation $\boldsymbol{y}_0=\boldsymbol{A}\boldsymbol{\alpha}$ is often under-determined in face recognition, the solution to $\boldsymbol{\alpha}$ is not unique. The aim of SRC is to investigate the discriminative information from the reconstruction coefficients. To ensure that the nonzero entries of the solution are constrained in a certain class, SRC exploits the possibility of imposing a penalty of l_1 norm regularization to the coefficients (Eq. (2-9)). Since l_1 norm minimization constrains a small absolute value of the solution, it is believed that the test image can be well represented with only a few nonzero entries.

The algorithm of SRC is briefly described below:

1. Input the training sample $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_k]$ with k classes and the test sample $\boldsymbol{y}_0$. The columns of $\boldsymbol{A}$ are normalized to have the unit l_2-norm.

2. Choose a suitable λ and solve the l_1-minimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}}\{\|\boldsymbol{y}_0 - \boldsymbol{A}\boldsymbol{\alpha}\|_2^2 + \boldsymbol{\lambda}\|\boldsymbol{\alpha}\|_1\} \tag{2-11}$$

3. Compute the residuals

$$\boldsymbol{e}(i) = \|\boldsymbol{y}_0 - \boldsymbol{A}_i\hat{\boldsymbol{\alpha}}_i\|_2^2 \quad \text{for} \quad i = 1,2,...,k \tag{2-12}$$

Where $\hat{\boldsymbol{\alpha}}_i$ is the sub-coefficients associated with the *i*th in $\hat{\boldsymbol{\alpha}}$.

4. Output that identity($\boldsymbol{y}$) = argmin$_{\boldsymbol{i}}$ e($\boldsymbol{i}$).

2.3.2 Advantage and Limitation of SRC

SRC introduces the concept of sparsity into face recognition. It is a powerful face classifier and achieves promising results in various face recognition tasks. Experimental results also show that SRC can handle the illumination variation and expression change, and also has discriminative ability when dealing with occlusion. However, the SRC model is based on the assumption that there are sufficient training samples in each class so that the test image can be well represented with the samples in the class where it belongs; actually, this is not always the cases in a face recognition system. There may not be enough samples in each class and they cannot represent the test image sufficiently, leading to the failure of SRC. That is the reason why SRC cannot handle the small sample size problem well, which is illustrated in the experiment part of Chapter 4.

It is important that SRC can ensure that the representation vector of the test sample $\boldsymbol{y_0}$ is sparse. To obtain such a stable sparse coefficient vector, considerable computation is needed to solve Eq. (2-9). The time consumed to reconstruct the test image under l_1 norm constraint makes implementation of SRC impossible in a real-time system. In addition, when SRC deals with face corruption or occlusion, an identity matrix $\boldsymbol{I}$ is introduced as the dictionary to encode the outlier points and thus the objective function becomes:

$$\min \left\| [\boldsymbol{\alpha}; \boldsymbol{\beta}] \right\|_1 \quad \text{s.t. } \boldsymbol{y} = [\boldsymbol{D}, \boldsymbol{I}] * [\boldsymbol{\alpha}; \boldsymbol{\beta}] \tag{2-13}$$

The identity matrix increases the computation cost greatly.

2.3.3 Collaborative Representation with Regularized Least Square (CR_RLS)

[15] does not investigate why SRC produces powerful recognition results while most previous works [62-65] focused on sparsity. However, [35] argues that the reason for SRC's promising performance is the collaborative representation of the test image. In Eq.

(2-9) it is assumed that the training samples of each class are sufficient to represent the test sample $\boldsymbol{y}_0$ and $\boldsymbol{A}_i$ is over-complete. However, it does not always hold true since $\boldsymbol{A}_i$ is under-complete generally, which means that even if $\boldsymbol{y}_0$ comes from the class i, the reconstruction error using $\boldsymbol{A}_i$ to code $\boldsymbol{y}_0$ may be still large. Therefore, more samples from other classes are needed to represent $\boldsymbol{y}_0$ when samples from the same class are insufficient. It is proved in [35] that when judging whether $\boldsymbol{y}$ belongs to a certain class i, the collaborative representation not only considers the distance between the ith class and $\boldsymbol{y}$, which is assumed to be small, but also the distance between y and other classes, which is assumed to be large. It is this "double consideration" that makes the collaborative representation, or SRC, robust and effective.

Since the l_1-norm sparse constraint does not account for FR results, the regularized least square constraint, also known as l_2-norm constraint, is proposed to replace the l_1-norn regularization to significantly reduce computation complexity while recognition performance is still effective. That is

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \{ \| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\alpha} \|_2^2 + \lambda \| \boldsymbol{\alpha} \|_2^2 \} \tag{2-14}$$

where λ is the regularization parameter. Here λ is mainly for making the solution stable instead of introducing sparsity to the solution. It is believed in [35] that the "sparsity" of coefficients associated with each class contains discriminative information. As a result, the coefficients are introduced into the determination of the identity, e.g.,

$$\boldsymbol{e}(i) = \| \boldsymbol{y}_0 - \boldsymbol{A}_i \hat{\boldsymbol{\alpha}}_i \|_2^2 / \| \hat{\boldsymbol{\alpha}}_i \|_2^2 \tag{2-15}$$

A brief description of the CR_RLS algorithm is shown below:

1. Input the training sample $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_k]$ with k classes and normalized into unit l_2-norm and the test sample $\boldsymbol{y}_0$.

2. Choose a suitable λ and solve the l_2-minimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \{ \| \boldsymbol{y}_0 - \boldsymbol{A}\boldsymbol{\alpha} \|_2^2 + \lambda \| \boldsymbol{\alpha} \|_2^2 \} \tag{2-16}$$

so that

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{A}\boldsymbol{A}^T + \lambda \boldsymbol{I})^{-1} * \boldsymbol{y}_0 = \boldsymbol{P} * \boldsymbol{y}_0 \tag{2-17}$$

3. Compute the residuals

$$\boldsymbol{e}(i) = \left\| \boldsymbol{y}_0 - \boldsymbol{A}_i \hat{\boldsymbol{\alpha}}_i \right\|_2^2 / \left\| \hat{\boldsymbol{\alpha}}_i \right\|_2^2 \quad \text{for} \quad i = 1, 2, ..., \mathrm{k} \tag{2-18}$$

where $\hat{\boldsymbol{\alpha}}_i$ is the sub-coefficients associated with the *i*th in $\hat{\boldsymbol{\alpha}}$.

4. Output identity($\boldsymbol{y}$) = argmin$_{\boldsymbol{i}}$ $\boldsymbol{e}$($\boldsymbol{i}$).

2.3.4 Advantage and Limitation of CR_RLS

In the CR_RLS model, the l_1-minimization constraint is removed and replaced with an l_2-minimization constraint, which can be easily solved by calculating $\boldsymbol{P} = (\boldsymbol{A}\boldsymbol{A}^T + \lambda \boldsymbol{I})^{-1}$. Since $\boldsymbol{P}$ can be obtained off-line, the only computation needed is to calculate the coefficients and reconstruction residual, and computation complexity of collaborative representation can be very low. Experimental results show that the performance of CR_RLS is similar to SRC, so it can also handle expression changes and illumination variations well. However, like SRC, CR_RLS cannot handle the small sample size problem well. In the next chapter, the l_2-minimization constraint is utilized in CR_RLS for designing our model, and CR_RLS, as a competing method, is competed and verified in popular databases.

Chapter 3. Fast Robust Face Recognition via Coding Residual Map Learning based Adaptive Masking (CRMLAM)

The face occlusion problem is one of the most challenging problems in FR. Recent sparse representation based classification (SRC) [15] and other sparse-based schemes could handle some face occlusion problems, however, their performance in clean faces is more remarkable than handling occluded face problem. In this thesis, we propose a simple but effective scheme to handle the occlusion FR problem, namely CRMLAM, which utilizes the reconstruction error distribution feature to detect the outlier points (occlusion, disguise or other points that will obstruct the recognition). Extensive experiments on representative databases show the results are very competitive and have significant less complexity than other sparse-based FR schemes.

3.1 Dictionary Learning

Many previous works [65-71] indicate that a dictionary learnt from original training samples usually brings in a more robust recognition result. Recently it attracted many researchers' attention that learning an over-complete dictionary from natural images and then using such a dictionary for image analysis can obtain a more satisfactory result. In face recognition, the original image samples usually have much redundancy as well as noise and trivial information that can be negative to the recognition. In addition, the atoms in the dictionary learnt from the original face images have more general information, which is helpful for reconstruction and recognition.

In representation based FR, the recognition is achieved by coding the query image over a dictionary. One straightforward way is to use the original training samples as the dictionary, such as in SRC [15] and CRC_RLS [35]. Since the training samples are generally non-occluded face images, the occluded pixels in a query image usually cannot

be well reconstructed by the non-occluded samples and thus a larger proportion of occluded pixels have large errors than non-occluded ones, we can use the coding residual to detect the occluded pixels, for example by setting a detection threshold, and then masking the detected occluded pixels from face coding to achieve robust FR.

However, the face image has various structures, e.g., eyes, nose, mouth and cheek, which will have different variances of coding residuals. It is hard to use a global threshold to effectively detect the occlusions in different facial areas. In order to make the occlusion detection more accurate, it is expected that we can know the coding residual variances of different facial features so that the spatially adaptive occlusion detection can be achieved.

To the above end, we can learn a dictionary from the training samples, and use this dictionary to code the training samples. The variances of coding residuals can then be computed to build the coding residual map. By coding a query image over this dictionary and with the learned coding residual map, the occlusion pixels can be adaptively detected. In addition, compared with using the original training samples as the naïve dictionary for face representation, dictionary learning can also bring advantages such as removing noise and trivial structures for more accurate face representation, as well as making the representation more discriminative.

Various dictionary learning methods have been proposed for image processing [79, 80, 82, 83] and pattern recognition [81, 84]. In [84], a Fisher discrimination dictionary learning (FDDL) method was proposed for sparse representation based image recognition. Inspired by FDDL and considering that the sparsity on the coding coefficients is not that important for FR [35], we propose a simpler dictionary learning model. Denote by $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_c]$ the dictionary to be learned, where $\boldsymbol{D}_i$ is the class-specified sub-dictionary associated with class i. The dictionary $\boldsymbol{D}$ is learned from the training dataset $\boldsymbol{A}$. In general, we require that each column of the dictionary $\boldsymbol{D}_i$ is a unit vector, and the number of atoms in $\boldsymbol{D}_i$ is no more than the number of training samples in $\boldsymbol{A}_i$.

We denote by $\boldsymbol{X}_i$ the coding coefficient matrix of $\boldsymbol{A}_i$ over $\boldsymbol{D}$, and $\boldsymbol{X}_i$ =[$\boldsymbol{X}_i^1$; …; $\boldsymbol{X}_i^j$; …; $\boldsymbol{X}_i^c$], where $\boldsymbol{X}_i^j$ is the coding matrix of $\boldsymbol{A}_i$ over the sub-dictionary $\boldsymbol{D}_j$. Inspired by [84], we propose to learn the dictionary $\boldsymbol{D}$ by optimizing the following objective function:

$$\boldsymbol{J}_{(\boldsymbol{D},\boldsymbol{X})} = \arg\min_{(\boldsymbol{D},\boldsymbol{X})} \sum_{i=1}^{c} \{\boldsymbol{R}_i(\boldsymbol{D}) + \lambda_1 \left\|\boldsymbol{X}_i\right\|_2^2 + \lambda_2 \left\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}_i\right\|_2^2\} \tag{3-1}$$

where

$$\boldsymbol{R}_i(\boldsymbol{D}) = \left\|\boldsymbol{A}_i - \boldsymbol{D}\boldsymbol{X}_i\right\|_2^2 + \left\|\boldsymbol{A}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\right\|_2^2 + \sum_{\substack{j=1 \\ j\neq i}}^{c} \left\|\boldsymbol{D}_j\boldsymbol{X}_i^j\right\|_2^2 \tag{3-2}$$

and $\bar{\boldsymbol{X}}_i$ is the column mean matrix of $\boldsymbol{X}_i$, i.e., every column of $\bar{\boldsymbol{X}}_i$ is the mean vector $\boldsymbol{m}_i$ of all the columns in $\boldsymbol{X}_i$. The parameters λ_1 and λ_2 are positive scalar numbers to balance the l_2-norm terms in Eq. (3-1).

From Eq. (3-2), one can see that the term $R_i(\boldsymbol{D})$ ensures that the training samples from class i (i.e., $\boldsymbol{A}_i$) can be well reconstructed by the learnt sub-dictionary $\boldsymbol{D}_i$, while they have small representations on the other sub-dictionaries $\boldsymbol{D}_j$, $j\neq i$. Therefore, the learned whole dictionary $\boldsymbol{D}$ will be discriminative in terms of reconstruction. On the other hand, the term $\left\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}_i\right\|_2^2$ in Eq. (3-1) will make the representations of the samples from the same class close to each other, reducing the intra-class variations. Finally, we use the l_2-norm, instead of the l_1-norm, to regularize the coding coefficients $\boldsymbol{X}$ in Eq. (3-1), and this significantly reduces the complexity of optimizing Eq. (3-1).

The objective function in Eq. (3-1) is a joint optimization problem of $\boldsymbol{D}$ and $\boldsymbol{X}$, and it is convex to $\boldsymbol{D}$ or $\boldsymbol{X}$ when the other is fixed. Like in many multi-variable optimization problems, we could solve Eq. (3-1) by optimizing $\boldsymbol{D}$ and $\boldsymbol{X}$ alternatively from some initialization. Since in each step, the optimization is convex and all the terms involved are of l_2-norm, the optimization can be easily accomplished. The learnt dictionary $\boldsymbol{D}$ will be different for different settings of parameters λ_1 and λ_2. Our experimental results show that

the final FR rates are not sensitive to λ_1 and λ_2 in a wide range. In our experiments, we set them as λ_1=0.001 and λ_2=0.002 for all the databases by experience.

3.2 Residual Map Learning

It is known that when a face image is coded over a dictionary, the residual $\boldsymbol{e}=\boldsymbol{D}-\boldsymbol{y}\boldsymbol{\alpha}$ usually follows the Gaussian or Laplacian distribution. However, when there are occlusion, corruption and/or other variance, the residual may be far from such distribution, since the occlusion or corruption distorts the original image structure, as shown in Fig. 3.1. Hence, the outlier residuals may indicate the corresponding original position is an occlusion or corruption. In fact, it is possible to "train" a variance map indicating the distribution of the coding residuals of the training set, which is supposed to be "clean" faces images without any occlusion. Such a variance map is able to act as the benchmark value of the normal residual in a certain position.

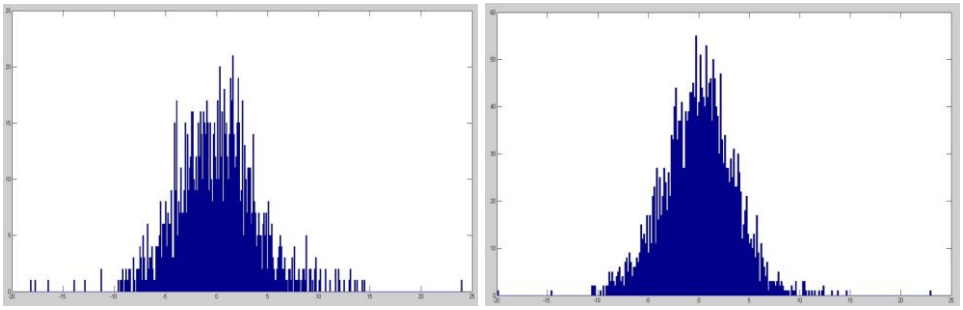


Figure 3.1: Histogram of coding residual of an occluded face image (left one) and a non occluded face image (right one)

Once the dictionary $\boldsymbol{D}$ is learnt from the training dataset $\boldsymbol{A}$ through Eq. (3-1), it can be used to code a given query sample $\boldsymbol{y}$ by

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha}\{\|\boldsymbol{y}-\boldsymbol{D}\boldsymbol{\alpha}\|_2^2+\gamma\|\boldsymbol{\alpha}\|_2^2\} \tag{3-3}$$

The solution $\hat{\boldsymbol{\alpha}}$ can be easily calculated as $\hat{\boldsymbol{\alpha}}=\boldsymbol{P}*\boldsymbol{y}$, where the projection matrix $\boldsymbol{P}=(\boldsymbol{D}^T\boldsymbol{D}+\gamma*\boldsymbol{I})^{-1}\boldsymbol{D}^T$ can be pre-computed. Then we can calculate the coding residual $\boldsymbol{e}_y=\boldsymbol{y}-\boldsymbol{D}\hat{\boldsymbol{\alpha}}$. When there are occlusions in the query image $\boldsymbol{y}$, its coding residuals

at the occluded pixels in $\boldsymbol{e}_y$ will probably exceed the "normal range" of $\boldsymbol{e}_y$. Therefore, if we could know the "normal range", or more specifically the standard deviation, of each element of $\boldsymbol{e}_y$, we can then detect adaptively the occlusions in $\boldsymbol{y}$.

Obviously, the deviation of the coding residuals will vary with the different facial areas. The areas such as eyes and mouth will have larger residuals than the areas such as cheek because they have more edge structures which are more difficult to reconstruct. This spatially varying coding residual deviation map can be learnt by coding the training samples in $\boldsymbol{A}$ over the dictionary $\boldsymbol{D}$. There is

$$\hat{\boldsymbol{\Lambda}} = \arg\min_{\boldsymbol{\Lambda}} \{ \|\boldsymbol{A} - \boldsymbol{D}\boldsymbol{\Lambda}\|_2^2 + \gamma \|\boldsymbol{\Lambda}\|_2^2 \} \tag{3-4}$$

Clearly, $\hat{\boldsymbol{\Lambda}} = \boldsymbol{P} \cdot \boldsymbol{A}$ and the coding residual matrix is $\boldsymbol{E} = \boldsymbol{A} - \boldsymbol{D}\hat{\boldsymbol{\Lambda}}$.

Each row of $\boldsymbol{E}$, denoted by $\boldsymbol{e}_k$, contains the coding residuals of all face samples at the same location k, and thus its standard deviation can be used to define the normal range of the coding residual at this location. Denoted by

$$\sigma_k = std(\boldsymbol{e}_k) \tag{3-5}$$

the standard deviation of $\boldsymbol{e}_k$, and then all the σ_k together will build a coding residual map, which indicates the normal range of coding residual at each location. Certainly, from Eq. (3-5) we know that the residual map depends on the parameter γ, and Fig. 3.1 shows several residual maps calculated on the AR database [72] with different values of γ. It can be seen that the different face structures will have different residual deviation, while the residual map varies with γ. Therefore, a suitable γ must be determined for robust FR.

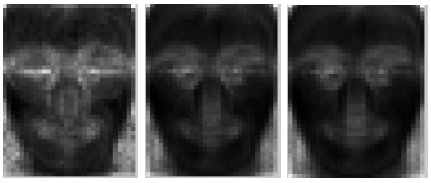


Figure 3.2: Examples of the learnt coding residual map with different settings of γ. From left to right, γ=0.1, 1, 2, respectively.

We determine γ by checking which value of γ can make the coding in Eq. (3-4) optimal. It can be empirically found that the coding residuals in $\boldsymbol{E}$ and the coding coefficients in $\boldsymbol{\Lambda}$ are nearly Gaussian distributed. We assume that the residuals in $\boldsymbol{E}$ and the coefficients in $\boldsymbol{\Lambda}$ follow i.i.d. Gaussian distributions, respectively. Based on the *maximum a posterior* (MAP) principle, the desired coefficients $\boldsymbol{\Lambda}$ should make the probability $P(\boldsymbol{\Lambda}|\boldsymbol{A})$ maximized. According to the Bayesian formula and after some straightforward derivations, the parameter γ which could lead to the MAP solution of $\boldsymbol{\Lambda}$ should satisfy

$$\gamma_{opt} = \|\boldsymbol{E}\|_2^2 \big/ \|\boldsymbol{\Lambda}\|_2^2 \tag{3-6}$$

In implementation, we use a set of different values of γ to code $\boldsymbol{A}$ by Eq. (3-5), thus the corresponding coefficient $\boldsymbol{\Lambda}$ and residual $\boldsymbol{E}$ can be obtained. Then we could check if the used γ is close enough to the associated γ_{opt} in Eq. (3-6). The γ which is the most close to its associated γ_{opt} is then used to compute the residual map elements σ_k.

3.3 Detecting the Occlusion Pixels

Once the residual map is learnt, we can use it to detect the occluded outlier pixels in query image $\boldsymbol{y}$ based on its coding residual $\boldsymbol{e}_y$. Intuitively, at location k, if $|\boldsymbol{e}_y(k)|$ is bigger enough than σ_k in the residual map, we could say that pixel k in $\boldsymbol{y}$ is occluded. It can be empirically found that the coding residual at location k, i.e., $\boldsymbol{e}_k$, approximately follows zero-mean Gaussian distribution, while the shape of the Gaussian distribution is controlled by σ_k. Fig. 3.2 plots the histograms of $\boldsymbol{e}_k$ at three different types of areas, eye, nose and cheek,

respectively. We can see that the distributions are Gaussian like, and the eye and nose regions have much higher standard deviation values than the smooth cheek area.

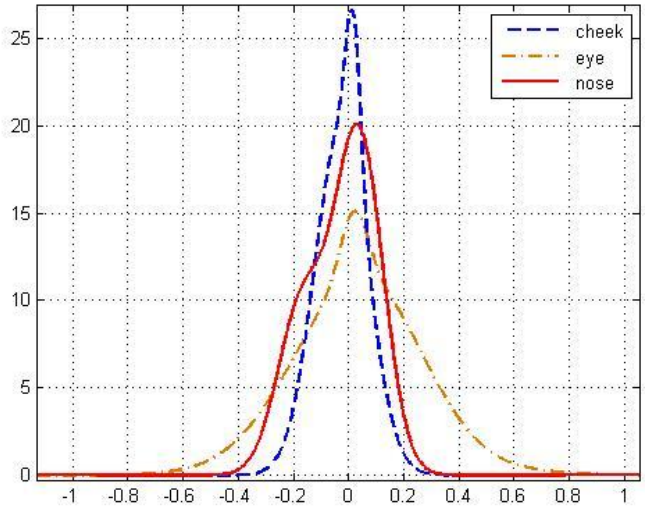


Figure 3.3: Histograms of the coding residuals at regions of eye (in brown), nose (in red) and cheek (in blue), respectively (Y axis: Frequency, X axis: Reconstruction Error).

It is known that most of the values of a Gaussian distribution will fall into the interval bounded by several times of its standard deviation. For example, 68% of the coding residual errors are within |σ| region of the distribution, 95% of the coding residual errors are within |2*σ| region of the distribution; Therefore, if a pixel at location *k* is occluded, the coding residual at this location will often exceed the normal range, and $|\boldsymbol{e}_y(k)|>c\sigma_k$ is very likely to happen, where *c* is a constant.

In this study, we use the following simple rule

$$\textit{pixel k is occluded} \text{ if } |\boldsymbol{e}_y(k)|>c*\boldsymbol{\sigma}_k \tag{3-7}$$

to detect the occlusions (see Fig. 3.3).

Figure 3.4: Examples of face disguise and occlusion.

By comparing the residual of the test image with the variance map, it is possible that those outlier points can be located through a few simple steps.

Let $\boldsymbol{y}$ denote a testing image, coding over the dictionary learnt in the above algorithm, we have:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \{ \| \boldsymbol{y} - \boldsymbol{D}\boldsymbol{\alpha} \|_2^2 + \gamma \| \boldsymbol{\alpha} \|_2^2 \} \tag{3-8}$$

And the reconstruction residual r is computed as

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{D}\hat{\boldsymbol{\alpha}} \tag{3-9}$$

If the testing image contains some disguise or occlusion, the distribution of residual r in those occluded parts will be far away from the Gaussian or Laplacian distribution. From our above analysis, if a certain $\boldsymbol{r(i)}$ satisfies Eq. (3-7), it is highly possible that the ith point is an outlier.

3.4 Masking and Coding

After detecting the occluded pixels in $\boldsymbol{y}$, we can partition the query image $\boldsymbol{y}$ into two parts: $\boldsymbol{y}=[\boldsymbol{y}_{nc}; \boldsymbol{y}_{oc}]$, where $\boldsymbol{y}_{nc}$ denotes the non-occluded part and $\boldsymbol{y}_{oc}$ denotes the occluded part. Since each pixel in $\boldsymbol{y}$ has a corresponding row in the learnt dictionary $\boldsymbol{D}$, we can accordingly partition the dictionary $\boldsymbol{D}$ into two parts, i.e., $\boldsymbol{D}=[\boldsymbol{D}_{nc}; \boldsymbol{D}_{oc}]$, where $\boldsymbol{D}_{nc}$ is the sub-dictionary for $\boldsymbol{y}_{nc}$ and $\boldsymbol{D}_{oc}$ is for $\boldsymbol{y}_{oc}$.

Since the occlusion in the query image will deteriorate the FR accuracy, obviously we can exclude $\boldsymbol{y}_{oc}$ from coding and use only $\boldsymbol{y}_{nc}$ to recognize the identity of $\boldsymbol{y}$. Therefore, the coding after masking is performed as:

$$\hat{\boldsymbol{\alpha}}_{nc} = \arg\min_{\boldsymbol{\alpha}} \left\{ \left\| \boldsymbol{y}_{nc} - \boldsymbol{D}_{nc}\boldsymbol{\alpha} \right\|_2^2 + \gamma \left\| \boldsymbol{\alpha} \right\|_2^2 \right\} \tag{3-10}$$

The solution of Eq. (3-10) is $\hat{\boldsymbol{\alpha}}_{nc} = \boldsymbol{P}_{nc} \cdot \boldsymbol{y}_{nc}$ with $\boldsymbol{P}_{nc} = \left(\boldsymbol{D}_{nc}^T \boldsymbol{D}_{nc} + \gamma \cdot \boldsymbol{I} \right)^{-1} \boldsymbol{D}_{nc}^T$. Since $\boldsymbol{P}_{nc}$ depends on the input query sample $\boldsymbol{y}$ and it cannot be pre-computed, the calculation of $\hat{\boldsymbol{\alpha}}_{nc}$ is the most time-consuming step of our proposed scheme. Fortunately, we do not need to calculate $\boldsymbol{P}_{nc}$ explicitly, and we can use the efficient Conjugate Gradient Method (CGM) [75] to solve $\hat{\boldsymbol{\alpha}}_{nc}$. The detailed complexity analysis will be made in Section 3.6.5, and the running time comparison will demonstrate that the proposed scheme is much faster than state-of-the-art robust FR methods but with competitive FR accuracy.

3.5 Classification

After $\hat{\boldsymbol{\alpha}}_{nc}$ is obtained by solving Eq. (3-10), the coding coefficient and class-specific coding residual can be used to determine the identity of query image $\boldsymbol{y}$. The class-specific coding residual can be calculated as $e_i = \left\| \boldsymbol{y}_{nc} - \boldsymbol{D}_{nc\_i}\hat{\boldsymbol{\alpha}}_{nc\_i} \right\|_2^2$, where $\boldsymbol{D}_{nc\_i}$ and $\hat{\boldsymbol{\alpha}}_{nc\_i}$ are the sub-dictionary and sub-coding vector associated with class i, respectively. Recall that in the dictionary learning stage in Section 3.1, we have also learnt the mean coding vector $\boldsymbol{m}_i$ of each class. Denote by $\boldsymbol{m}_{nc\_i}$ the corresponding mean coding vector to the non-occluded face vector $\boldsymbol{y}_{nc}$. The distance between $\hat{\boldsymbol{\alpha}}_{nc\_i}$ and $\boldsymbol{m}_{nc\_i}$ can also help classifying $\boldsymbol{y}$. Let $g_i = \left\| \hat{\boldsymbol{\alpha}}_{nc} - \boldsymbol{m}_{nc\_i} \right\|_2^2$. Finally, we could fuse e_i and g_i for decision making:

$$f_i = e_i + w \cdot g_i \tag{3-11}$$

where weight w is a constant. The identity of the query image is then determined by: identity($\boldsymbol{y}$)=argmin$_i${f_i}.

The detail procedures of the CRMLAM algorithm are described as below:

The CRMLAM Algorithm

1. Normalize the columns of $\boldsymbol{A}$ to have unit l_2-norm.

2. Learn a dictionary $\boldsymbol{D}$ through Eq.(3-1) and Eq. (3-2)

3. Code $\boldsymbol{A}$ over the $\boldsymbol{D}$ via l_2-norm minimization

$$\hat{\boldsymbol{\Lambda}} = \arg\min_{\boldsymbol{\Lambda}} \{\|\boldsymbol{A} - \boldsymbol{D}\boldsymbol{\Lambda}\|_2^2 + \gamma |\boldsymbol{\Lambda}|_2^2\} \tag{3-12}$$

4. Construct the variance map

   Compute the reconstruction residual $\boldsymbol{E}$=$\boldsymbol{A}$-$\boldsymbol{D}\boldsymbol{\Lambda}$, calculate the variance row by row, that is $\boldsymbol{\sigma}(i)$=var(($\boldsymbol{E}(i,:)$)), i=1,2,…,n.

5. Code the test sample $\boldsymbol{y}$ over the dictionary, and then calculate the residuals

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{\alpha}\|_2^2 + \gamma \|\boldsymbol{\alpha}\|_2^2\} \tag{3-13}$$

$$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{D}\hat{\boldsymbol{\alpha}} \tag{3-14}$$

6. Detect the outlier pixels by checking whether the residuals satisfy $|\boldsymbol{r}(i)|>c*\boldsymbol{\sigma}(i)$, and discard those outlier points from the test sample as well as the dictionary accordingly to form the subset $\boldsymbol{y}_{org}$ and $\boldsymbol{D}_{org}$.

7. Code the $\boldsymbol{y}_{org}$ over $\boldsymbol{D}_{org}$ via l_2-norm minimization as

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \{\|\boldsymbol{y}_{org} - \boldsymbol{D}_{org}\boldsymbol{\alpha}\|_2^2 + \gamma |\boldsymbol{\alpha}|_2^2\} \tag{3-15}$$

   Then calculate the residual

$$e_i(\boldsymbol{y}_{org}) = \|\boldsymbol{y}_{org} - \boldsymbol{D}_{org}(i)\hat{\boldsymbol{\alpha}}_i\|_2^2 + \boldsymbol{w}*\boldsymbol{g}_i \tag{3-16}$$

   where

$\hat{\boldsymbol{\alpha}}_i$ is the coefficient vector and $\boldsymbol{D}_{org}(i)$ is the dictionary associated with class i.

8. Output the identity of $\boldsymbol{y}$ as

$$\text{identity}(\boldsymbol{y}) = \arg\min_{i}(\boldsymbol{e}_i) \tag{3-17}$$

3.6 Experiment Verification

In this section, we perform extensive experiments on benchmark face databases to demonstrate the performance of CRMLAM algorithm. We first discuss the parameters selection in Section 3.6.1; in Section 3.6.2 we test the proposal algorithm on three databases (AR [72], Extended YaleB [56], Mpie [73]) without disguise and occlusion; in Section 3.6.3, different types of disguise and occlusion will be added into the AR and Extended YaleB data to test the performance of the algorithm when handling outliers. Finally, we will discuss the computation efficiency in Section 3.6.4.

3.6.1 Parameters Selection

There are five parameters in our algorithm: λ_1 and λ_2 in Eq. (3-1), γ in Eq. (3-3), (3-4), the constant c in Eq. (3-7) and the weight w in Eq. (3-11). Under l_2 norm constraint, the regularized parameters are no longer critical for recognition results [35], so λ_1 and λ_2 are not sensitive to experiments results and they are set as 0.001 and 0.002 by experience, respectively, for all datasets. The value of γ is determined through Eq. (3-6). The weight w can be set empirically and it is fixed as 0.1 in all experiments.

The value of c is critical to detect the occlusion points. If c is too small, too many points are deemed as outliers; if c is too large, few outliers will be detected. Fig. 3.4 shows some detection examples. By experience, in our following experiments, c is set as 2, 1 and 1 in the AR, Extended Yale B and MPIE databases, respectively. In fact, when c=1~2, the recognition results will not change a lot (less than 2%). If the value of c is set in this range, the recognition result is still acceptable; the parameters in other competing methods, e.g. the λ in CRC and SRC, also needs to adjust, so we think our comparison is still fair.

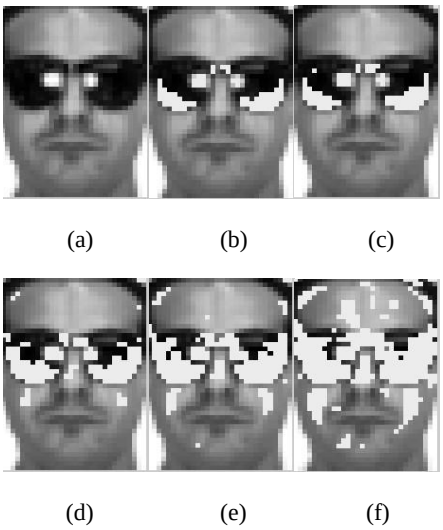


Figure 3.5: Example of occlusion point detection. (a) is the original test image. From (b) to (f): the occlusion detection results by letting c=12, 10, 6, 4, 2, respectively.

3.6.2 Recognition without Occlusion

Although our algorithm mainly focuses on handling occlusion, in clean face images without occlusion or disguise, there are still some pixels that can be viewed as outliers which may lead to recognition error. We are trying to detect these pixels with our proposed method and exclude them from recognition. We evaluate our algorithm in the popular databases AR [72], Extend Yale B [56] and MPIE [73].

a) AR database: The AR database [72] consists of over 4,000 frontal images from 126 individuals. For each individual, 26 pictures were taken in two separate sessions. The setting in our experiments is the same as in [15]. A subset that contains 50 males and 50 females with only illumination and expression variances is used. For each individual, the seven images from Section 1 are used as training samples while other seven images from Section 2 are used as testing. The size of original face image is downsampled to 36×22 in our method. The best results of competing methods, including nearest neighbor (NN) classifier, SRC [15], CRC_RLS [35], RSC [49] and the CRMLAM method, are presented in Table 3.1.

CRC_RLS and the CRMLAM method have similar reconstruction strategy, while the CRMLAM achieves higher recognition rate. This demonstrates that the proposed outlier pixel detection method can remove some insignificant (even negative) pixels in the face images, and hence improve the recognition rate. Compared with NN and SRC, the proposed CRMLAM also achieves about 23.8% and 1.8% higher recognition rate. The CRMLAM algorithm achieves the second best accuracy and is only slightly worse than RSC, whose complexity is much higher than our method (please refer to Section 3.6.5 for the running time comparison).

Table 3.1: Recognition rates on the AR database by different methods.

NN	SRC	CRC_RLS	RSC	CRMLAM
71.3%	93.3%	93.7%	96.0%	95.1%

b) Extended Yale B Database: The extended Yale B [56] database contains about 2,414 frontal face images of 38 individuals taken under varying illumination conditions. In our experiment, one half, which contains 32 images per person, is used as the training set, while another half, which also contains 32 images per person, is used as the testing set. All the images are cropped and down-sampled to 27×24 in the experiments. The comparison of different methods is shown in Table 3.2. Since the extended Yale B database has only 38 subjects and each subject has many training samples, all the methods except for NN could achieve the same good results.

Table 3.2: Recognition rates on the Extended Yale B database by different methods.

NN	SRC	CRC_RLS	RSC	CRMLAM
91.6%	98.3%	98.3%	98.3%	98.3%

c) Multi PIE database: The CMU Multi-PIE database [73] contains image of 337 subjects captured in four sessions with simultaneous variations in pose, expression, and illumination. Among these 337 subjects, all the 249 subjects in Session 1 are used. We use 7 frontal images with extreme illuminations {0, 1, 7, 13, 14, 16, 18} and neutral expression

of each subject as the training set. For testing set, 4 images of various illuminations taken with smile expressions each person in the same session are used. The face images used in this experiment are directly down-sampled to 25×20. Table 3.3 lists the recognition rates of different methods. Again, the proposed CRMLAM achieves the second best rate after the RSC scheme.

Table 3.3: Recognition rates on the MPIE database by different methods.

NN	SRC	CR_RLS	RSC	CRMLAM
86.4%	93.9%	94.1%	97.8%	94.4%

3.6.3 Recognition with Real Disguise

As in [15], a subset from the AR database consists of 1,400 images from 100 subjects, 50 male and 50 female, is used here. 800 images (about 8 samples per subject) of non-occluded frontal views with various facial expressions in both sessions are used for training, while the samples with sunglasses (2 samples per subject) and scarves (1 sample per subject) in both sessions are used for testing. Fig. 3.5 shows some example query images with disguise. The images are directly down-sampled to 42×30 with normalization.

Figure 3.6: Testing samples with sunglasses and scarves in the AR database.

The occlusion pixels detected by the proposed CRMLAM are illustrated in Fig. 3.6. It can be seen that the CRMLAM algorithm detects many outlier points, and also makes some wrong judgments. Fortunately, the number of falsely detected outliers is not big so

that the recognition will not be much affected. For smaller ratio of occlusion (e.g., sunglasses disguise), the detection result is more accurate (as shown in Figs. 3.6(a) and 3.6(c)); but when the occlusion ratio is large (e.g., the scarf disguise), the detection result is less accurate (as shown in Figs. 3.6(b) and 3.6(d)).

The recognition rates by competing methods, including NN, SRC, CRC_RLS, RSC and the proposed CRMLAM, are listed in Table 3.4. Although the occlusion detection is not accurate enough, the proposed scheme can still obtain much better results than all the competing methods except for RSC. Again, we would like to emphasize that RSC has much higher complexity than the proposed CRMLAM.

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

Figure 3.7: The occlusion detection results of some testing samples in the AR database.

Table 3.4: Recognition results by different methods on the AR database with sunglasses and scarves disguise.

Algorithm	Sunglass	Scarve
NN	70.0%	12.0%
SRC	87.0%	59.5%
CRC_RLS	68.5%	90.5%
RSC	99.0%	97.0%
CRMLAM	93.0%	90.5%

3.6.4 Recognition with Random Block Occlusion

In this section, we test the robustness of our algorithm to random block occlusion. As in [15], Subset 1 and 2 of the Extended Yale B database are used for training and Subset 3 for testing. Each testing sample will be inserted an unrelated image as block occlusion, and the blocking ratio is from 10% to 50% as illustrated in Fig. 3.7. The images are cropped and down-sampled to 48×42.

All training and testing samples are normalized to reduce the effect of illumination variance. The occlusion detection results of the example images in Fig. 3.7 are shown in Fig. 3.8. The recognition rates by the competing methods are listed in Table 3.5. We can see that when the block occlusion ratio is low, all methods can achieve good recognition accuracy; when the block occlusion ratio increases, the accuracy of NN, CRC_RLS and SRC will decrease rapidly, while RSC and the proposed CRMLAM can still have good results. When the occlusion ratio is 50%, the proposed CRMLAM surpasses SRC more than 10%, while being only about 6% lower than RSC.

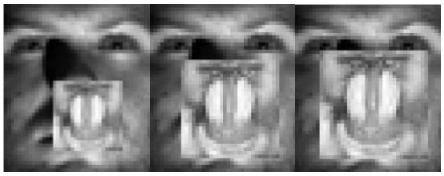


Figure 3.8: Examples of random block occlusion in the Extended Yale B database. From left to right: occlusion ratio is 20%, 40%, 50%, respectively.

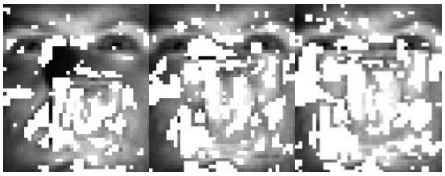


Figure 3.9: Occlusion detection results of the samples in Fig. 3.7.

Table 3.5: Recognition results by different methods on the Extended Yale B database with various random occlusion ratios.

Occlusion	10%	20%	30%	40%	50%
NN	90.1%	85.2%	74.2%	63.8%	48.1%
SRC	100%	99.8%	98.5%	90.3%	65.3%
CRC_RLS	99.8%	93.6%	82.6%	70.0%	52.3%
RSC	100%	100%	99.8%	96.9%	83.9%
CRMLA	100%	99.8%	98.5%	93.6%	77.9%

3.6.5 Complexity Analysis

From the experimental results in Sections 3.6.2~3.6.4, we see that the CRMLAM's recognition accuracy is only slightly lower than RSC but much higher than SRC and

CRC_RLS, which are among the state-of-the-arts. Let's then compare the time complexity and running time between our method and SRC, CRC_RLS and RSC.

The computational cost in our method mainly comes from solving Eq. (3-15). The solution of Eq. (3-15) is

$$\hat{\boldsymbol{\alpha}}_{nc} = \left(\boldsymbol{D}_{nc}^T \boldsymbol{D}_{nc} + \gamma \cdot \boldsymbol{I}\right)^{-1} \boldsymbol{D}_{nc}^T \cdot \boldsymbol{y}_{nc} \tag{3-18}$$

We can re-write it as

$$\boldsymbol{D}_{nc}^T \cdot \boldsymbol{y}_{nc} = \left(\boldsymbol{D}_{nc}^T \boldsymbol{D}_{nc} + \gamma \cdot \boldsymbol{I}\right) \hat{\boldsymbol{\alpha}}_{nc} \tag{3-19}$$

Eq. (3-19) can be viewed as a linear equation system, where the coefficient $\hat{\boldsymbol{\alpha}}_{nc}$ is to be computed.

Here we use the Conjugate Gradient Method (CGM) [75] algorithm to solve this linear equation system, which is much more efficient than solving the inverse problem in Eq. (3-18). Suppose that $\boldsymbol{D}_{nc}$ is an $N \times M$ matrix. In FR problems, usually we have $N > M$. The solution of Eq. (3-19) is a vector of length M. The computational complexity of the CGM is $O(M)$ for each iteration, so the complexity of solving Eq. (3-19) is $O(KM)$ if there are totally K iterations. Since we do not need a very accurate solution to Eq. (3-19), we set K=30 in our experiments as the maximum number of iteration. The complexity for calculating $\boldsymbol{D}_{nc}^T \boldsymbol{D}_{nc}$ is $O(NM^2)$. Thus the total time complexity of our method is $O(NM^2+KM)$.

The following Tables 3.6~3.8 show the running time of CRC_RLS [35], SRC with l_1_*ls* [76] and SRC with fast homotopy [77] (for other fast l_1-norm minimization methods please refer to [78]), RSC [49] and our proposed CRMLAM in three experiments, which are conducted under MATLAB environment on a PC with Intel R Core 2 1.86 GHz CPU and 2.99GB RAM. The settings are the same as those in previous sections, i.e., AR with sunglass disguise, Extended Yale B with 50% occlusion, and MPIE without occlusion. All

samples are directly down-sampled from the original face images. The reported running time is the average time consumed by each testing sample.

Table 3.6: Recognition rates and running time on the AR database with sunglass disguise.

Algorithm	Recognition Rate	Running
SRC (l_1_ls)	87.0%	2.403s
SRC (homotopy)	86.4%	0.120s
CRC_RLS	68.5%	0.017s
RSC	99.0%	160s
CRMLAM	93.0%	0.141s

Table 3.7: Recognition rates and running time on the Extended Yale B database with 50% block occlusion.

Algorithm	Recognition Rate	Running
SRC (l_1_ls)	65.3%	40.24s
SRC (homotopy)	63.5%	0.115s
CRC_RLS	52.3%	0.034s
RSC	83.9%	270s
CRMLAM	77.9%	0.278s

Table 3.8: Recognition rates and running time on the MPIE database without occlusion.

Algorithm	Recognition Rate	Running
SRC (l_1_ls)	93.9%	59.25s
SRC (homotopy)	92.0%	0.560s
CRC_RLS	94.1%	0.417s
RSC	97.8%	340s
CRMLAM	94.4%	0.858s

From Tables 3.6~3.8, we can make the following findings. First, in all the experiments, RSC always achieves the best recognition rates, while the proposed CRMLAM method always has the second best recognition rates. However, the speed of our method is hundreds to thousands of times faster than RSC. Second, CRC_RLS is the fastest algorithm among all the competing methods. It has similar recognition rate to our proposed CRMLAM method for FR without occlusion (please refer to Table 3.8), but has much worse recognition rates than ours for FR with occlusion. Third, SRC implemented by homotopy techniques has similar running time to our method, whereas its recognition rates

are lower than our method, especially for FR with occlusion. Finally, SRC implemented by l_1_ls is very slow without improving much the recognition rates compared to SRC implemented by homotopy. In summary, the RSC is so far one of the algorithms that achieve the best result in occlusion face recognition. The CRMLAM aims to find a very fast scheme for face occlusion removal, which can be naturally embedded into a coding based FR framework. The RSC scheme is very accurate for occlusion detection but it is thousands of times slower than CRMLAM, while CRMLAM is only 1%~6% lower than RSC in FR rate in various tests, and we think the CRMLAM model can well balance the computation cost and recognition accuracy. The RSC here is used as a benchmark to evaluate the CRMLAM. the proposed CRMLAM achieves a very good balance between robustness and efficiency. In practical FR applications, the database can be of large scale, and our method could lead to desirable recognition accuracy with acceptable time consumption.

3.7 Conclusion

In this chapter, we proposed a simple yet robust and efficient FR scheme by coding the query sample over a dictionary learnt from the training samples. To make the FR robust to occlusions and disguise, a coding residual map was first learnt from the training samples, and then it is used to adaptively detect the outlier points in the query sample. The detected outliers are then excluded from the coding of the query sample to improve the robustness of FR with occluded samples. Our extensive experimental results on benchmark face databases show that the proposed scheme is very competitive to state-of-the-arts in terms of accuracy, while it is much faster than the methods such as SRC [15] and RSC [49]. Overall, the proposed method has both robust FR performance and fast speed. It is a very good candidate for robust and real-time face recognition applications.

Chapter 4. A Joint Discriminative Dimensionality Reduction and Dictionary Learning (JDDRDL) Algorithm for Face Recognition

In Chapter 3, we proposed a robust face recognition method to handle the occlusion problem. In this chapter, we will propose a novel and effective joint learning algorithm, namely JDDRDL to handle the dictionary learning and dimension reduction and it is robust to the small sample size problem. We define an objective function to learn a dictionary to represent the training samples as well as an optimal discriminative dimension reduction matrix to code and classify the testing samples in a much lower dimension space. Also, a classification scheme associated with the reconstruction error and the coefficients information is applied. The proposed algorithm is extensively evaluated on benchmark face databases in comparison with existing dictionary learning and dimension reduction methods.

4.1 Motivation to Propose the JDDRDL Method

There has been increasing interest in learning a dictionary to represent the query image (as we have illustrated in Chapter 3) instead of using the original training samples. In face recognition, the original face images may contain some redundant information, noise or other trivial information that will obstruct recognition. In [59], M. Yang *et al.* proposed a metaface learning (MFL) algorithm to represent the training samples, a series of "metaface" are learnt from each class with a sparse constraint. [80] proposed a K-SVD algorithm, in which an over-complete dictionary is learnt also on a sparse constraint. Such a dictionary could be used in image denoising, super-resolution, recognition. In [81], a supervised dictionary learning method is proposed and applied in sparse signal decompositions in image classification tasks (digit recognition, texture classification). [33] introduced a Gabor Occlusion Dictionary to reduce the computation complexity in

occlusion face recognition in the frame of SRC, and received much better result. [82] developed a class-dependent supervised simultaneous orthogonal matching pursuit scheme to solve the dictionary learning problem, which increases the inter-class discrimination. Very recently, a Fisher discrimination dictionary learning algorithm [84] was developed for sparse representation based pattern classification, and it shows very competitive performance with other dictionary learning based pattern classification schemes. In [86], Zhang *et al*. proposed a discriminative K-SVD scheme to incorporate the classification error into the objective function and bring in efficiency and performance increase. In [87], it is proposed to train the K-SVD dictionaries for predefined face image patches, and compress those images according to the dictionaries. More review about the dictionary learning sparse representation could be found in [83].

The dimensionality reduction (DR) and dictionary learning (DL) are mostly studied as two independent problems in FR. Usually, DR is performed first to the training samples and the dimensionality reduced data are used for DL. However, the pre-learnt DR projection may not preserve the best features for DL. Intuitively, the DR and DL processes should be jointly conducted for a more effective FR. To this end, we propose a joint discriminative DR and DL (JDDRDL) scheme to exploit more effectively and robustly the discriminative information of training samples. The goal is that the face image feature from different classes can be effectively separated by a dictionary in a subspace, which is to be determined. In the proposed JDDRDL, an energy function is defined and an iterative optimization algorithm is given to alternatively optimize the dictionary and projection matrix. From some initialization, in each iteration, for a fixed projection $\boldsymbol{P}$, the desired dictionary $\boldsymbol{D}$ can be updated; then with the updated dictionary $\boldsymbol{D}$, the projection matrix $\boldsymbol{P}$ can be refined. After several iterations, the learnt $\boldsymbol{P}$ and $\boldsymbol{D}$ together can lead to a more effective FR system.

One important advantage of the proposed JDDRDL scheme is that it is more robust to the small sample size problem than state-of-art linear representation based face

classification methods [15, 57, 59, 84]. The discriminative DR methods such as LDA and the linear representation based methods such as SRC usually require that the number of training samples per class cannot be too small, and their performance can be much reduced if the training sample is insufficient. By exploiting more effectively the discriminative information of training sample via learning the projection and dictionary simultaneously, the proposed JDDRDL shows more robust recognition capability for FR when the training sample size per class is small, for example 2~5 samples per class.

In next sessions, we will briefly review the relate representation based dimension reduction and dictionary learning method, and then present the experiment result.

4.2 DR and DL under the SRC Framework

It is claimed in [15] that SRC is insensitive to feature extraction when the dimensionality is high enough; however, a well learnt DR matrix can lead to a more accurate and stable recognition result. In [57], an orthogonal DR matrix $\boldsymbol{P}$ is learnt under the framework of sparse representation, and it achieves better performance than Eigenfaces and Randomfaces in the SRC scheme. Specifically, the matrix $\boldsymbol{P}$ is learnt via the following objective function based on Leave-One-Out scheme:

$$\boldsymbol{J}_{\boldsymbol{P},\{\boldsymbol{\beta}_i\}} = \arg\min_{\boldsymbol{P},\{\boldsymbol{\beta}_i\}} \{\sum_{i=1}^{N} (\left\|\boldsymbol{P}\boldsymbol{z}_i - \boldsymbol{P}\boldsymbol{A}_i\boldsymbol{\beta}_i\right\|_2^2 + \lambda_1 \left\|\boldsymbol{\beta}_i\right\|_1) + \lambda_2 \left\|\boldsymbol{A} - \boldsymbol{P}^T\boldsymbol{P}\boldsymbol{A}\right\|_2^2\} \text{ s.t. } \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I} \tag{4-1}$$

where N is the number of training samples, $\boldsymbol{z}_i$ is the ith sample of the training set $\boldsymbol{A}$, and $\boldsymbol{A}_i$ is the set of training samples in $\boldsymbol{A}$ excluding $\boldsymbol{z}_i$. As can be seen from the above objective function, the projection matrix $\boldsymbol{P}$ preserves the energy of training set $\boldsymbol{A}$ while keeping the coding vector of each sample $\boldsymbol{z}_i$ sparse.

In SRC, the original training samples are used as the dictionary to represent the query sample. Intuitively, a more accurate and discriminative representation can be obtained if we could optimize a dictionary from the original training samples. In [59], Yang *et al.*

proposed a "metaface" learning method, where a dictionary $\boldsymbol{D}_k = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p]$ of "metafaces" is learnt from each class of training samples $\boldsymbol{A}_k$ under the sparse representation model via optimizing $J_{\boldsymbol{D}_k, \boldsymbol{\Lambda}} = \arg\min_{\boldsymbol{D}_k, \boldsymbol{\Lambda}} \|\boldsymbol{A}_k - \boldsymbol{D}_k \boldsymbol{\Lambda}\|_2^2 + \lambda \|\boldsymbol{\Lambda}\|_1$ s.t. $\boldsymbol{d}_j^T \boldsymbol{d}_j = 1,\ j = 1, \ldots, p$. The metaface dictionary $\boldsymbol{D}_k$ and the associated coefficient matrix $\boldsymbol{\Lambda}$ are optimized alternatively. The final metaface dictionary $\boldsymbol{D}$ is formed by concatenating all the K dictionaries $\boldsymbol{D}_k$.

Though the metaface learning method [59] improves the representation power of the dictionary, it does not truly aim to increase the discrimination power of $\boldsymbol{D}$ in the objective function. Yang *et al.* [84] recently proposed a DL method, namely the Fisher discrimination dictionary learning (FDDL), which embeds the Fisher criterion in the objective function design. The FDDL scheme has two remarkable features. First, the dictionary atoms are learnt to associate the class labels so that the reconstruction residual from each class can be used in classification; second, the Fisher criterion is also imposed on the coding coefficients so that they carry discriminative information for classification. Since both the reconstruction residual and coding coefficients are discriminative, a new classification scheme is then proposed in FDDL to fuse the two types of information for a more robust pattern recognition task.

4.3 The JDDRDL Algorithm

4.3.1 JDDRDL Model

In the related works introduced in Section 2, we could see that the DR and DL processes are handled separately. Usually, the DR projection matrix can be learnt first to reduce the dimensionality of training samples, and then DL is performed to learn a dictionary from the dimensionality reduced dataset. To more effectively use the discrimination information in the training set $\boldsymbol{A}$, we propose to learn the DR matrix $\boldsymbol{P}$ and the dictionary $\boldsymbol{D}$ jointly so that a more accurate classification can be achieved.

For the projection matrix $\boldsymbol{P}$, we expect that it could preserve the energy of $\boldsymbol{A}$ while making the different classes $\boldsymbol{A}_i$ more separable in the subspace defined by $\boldsymbol{P}$. To this end, we will learn an orthogonal projection matrix, which could maximize the total scatter of $\boldsymbol{A}$ and the between-class scatter of $\boldsymbol{A}$ simultaneously. For the dictionary $\boldsymbol{D}$, we expect that it could be able to faithfully represent the dimensionality reduced dataset $\boldsymbol{PA}$, while making the samples from the same class close to each other in the space spanned by $\boldsymbol{D}$. With the above considerations, in this chapter we propose the following joint discriminative dimensionality reduction and dictionary learning (JDDRDL) model to optimize $\boldsymbol{P}$ and $\boldsymbol{D}$:

$$J_{\boldsymbol{P},\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}} = \underset{\boldsymbol{P},\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}}{\arg\min} \left\{ \begin{array}{l} \sum_{k=1}^{K} \left(\left\| \boldsymbol{PA}_k - \boldsymbol{D}_k\boldsymbol{\Lambda}_k \right\|_2^2 + \lambda_1 \left\| \boldsymbol{\Lambda}_k \right\|_2^2 + \lambda_2 \left\| \boldsymbol{\Lambda}_k - \boldsymbol{\Gamma}_k \right\|_2^2 \right) \\ -\gamma_1 \left\| \boldsymbol{PA}_t \right\|_2^2 - \gamma_2 \left\| \boldsymbol{PA}_b \right\|_2^2 \end{array} \right\} \text{ s.t. } \boldsymbol{d}_{k,j}^T \boldsymbol{d}_{k,j} = 1,\ \forall k, j,\ \boldsymbol{PP}^T = \boldsymbol{I} \quad (4\text{-}2)$$

where $\boldsymbol{D}_k$ is the sub-dictionary for class k and $\boldsymbol{D}=[\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_K]$ forms the whole dictionary; $\boldsymbol{\Lambda}_k$ represents the coding coefficient matrix of $\boldsymbol{PA}_k$ over $\boldsymbol{D}_k$; $\boldsymbol{A}_t$ is centralized training set, i.e., $\boldsymbol{A}_t = \boldsymbol{A}-\boldsymbol{M}$ with each column of $\boldsymbol{M}$ being the mean vector $\boldsymbol{m}$ of all samples in $\boldsymbol{A}$; $\boldsymbol{A}_b$ is the class-specific centralized dataset of $\boldsymbol{A}$, i.e., $\boldsymbol{A}_b = [\boldsymbol{M}_1-\boldsymbol{M}, \ldots, \boldsymbol{M}_K-\boldsymbol{M}]$ with each column of $\boldsymbol{M}_k$ being the mean vector $\boldsymbol{m}_k$ of samples in $\boldsymbol{A}_k$; $\boldsymbol{\Gamma}_k$ is a matrix with each column of it being the mean of the columns in $\boldsymbol{\Lambda}_k$; λ_1, λ_2, γ_1, and γ_2 are positive scalars. We require that each atom $\boldsymbol{d}_{k,j}$ in dictionary $\boldsymbol{D}_k$ has unit norm.

Let's make a more detailed look of the JDDRDL model in Eq. (4-2). By requiring that $\boldsymbol{P}$ is orthogonal, minimizing the term $-\left\| \boldsymbol{PA}_t \right\|_2^2$ (i.e., maximizing $\left\| \boldsymbol{PA}_t \right\|_2^2$) guarantees that the energy of $\boldsymbol{A}_t$ can be well preserved because we can reconstruct $\boldsymbol{A}_t$ by $\boldsymbol{P}^T\boldsymbol{PA}_t$. On the other hand, minimizing the term $-\left\| \boldsymbol{PA}_b \right\|_2^2$ will enhance the discrimination between different classes after projection because it aims to maximize the distance between the class centers. Minimizing $-\left\| \boldsymbol{PA}_t \right\|_2^2$ and $-\left\| \boldsymbol{PA}_b \right\|_2^2$ simultaneously will also make the within class scatter of dataset $\boldsymbol{A}$ small.

By coding $\boldsymbol{PA}_k$ over $\boldsymbol{D}_k$, we minimize the coding residual $\|\boldsymbol{PA}_k - \boldsymbol{D}_k\boldsymbol{A}_k\|_2^2$ to ensure the representation power of dictionary $\boldsymbol{D}_k$. Note that we use the l_2-norm, instead of the sparse l_1-norm, to regularize the coding coefficients by $\|\boldsymbol{\Lambda}_k\|_2^2$. This is based on the recent findings [35] that the l_1-norm sparsity does not play the key role in sparse representation based FR. However, using the l_2-norm to regularize $\boldsymbol{\Lambda}_k$ significantly reduces the time complexity for optimization without sacrificing the performance. Finally, the minimization of $\|\boldsymbol{\Lambda}_k - \boldsymbol{\Gamma}_k\|_2^2$ enforces the coding coefficients of the samples in class k to be close to their mean, reducing the variations of the coding vectors of each class. This minimizes the within class scatter but in the domain spanned by the dictionary $\boldsymbol{D}_k$.

Overall, in the JDDRDL model in Eq. (4-2), the targeted projection $\boldsymbol{P}$ and dictionary $\boldsymbol{D}$ will make the training samples have larger between class distance and smaller within class variation. Ideally, if $\boldsymbol{P}$ and $\boldsymbol{D}$ could be well optimized, more accurately classification of the query sample $\boldsymbol{y}$ can be obtained. Next, let's discuss how to do the minimization of Eq. (4-2).

4.3.2 The Optimization

The JDDRDL objective function in Eq. (4-2) is non-convex, and here we propose an alternative optimization algorithm in order for a locally optimal solution. We partition the whole optimization into two sub-problems: fix the projection matrix $\boldsymbol{P}$ and solve for the dictionary $\boldsymbol{D}$ and the coefficient $\boldsymbol{\Lambda}$; and fix $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$ to update $\boldsymbol{P}$. These two sub-problems are solved alternatively and iteratively, and we stop at a good point to get the locally optimal solutions of $\boldsymbol{P}$ and $\boldsymbol{D}$. The whole optimization algorithm is presented in detail as follows.

Step 1) Initialize $\boldsymbol{P}$. We use PCA to initialize $\boldsymbol{P}$. That is, the initial $\boldsymbol{P}$ is the PCA transformation matrix of the training data $\boldsymbol{A}$.

Step 2) Fix $\boldsymbol{P}$, and solve $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$. In this case, the objective function in Eq. (4-2) reduces to

$$J_{\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}} = \arg\min_{\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}} \sum_{k=1}^{K} \left(\left\| \boldsymbol{P}\boldsymbol{A}_k - \boldsymbol{D}_k\boldsymbol{\Lambda}_k \right\|_2^2 + \lambda_1 \left\| \boldsymbol{\Lambda}_k \right\|_2^2 + \lambda_2 \left\| \boldsymbol{\Lambda}_k - \boldsymbol{\Gamma}_k \right\|_2^2 \right) \text{ s.t. } \boldsymbol{d}_{k,j}^T \boldsymbol{d}_{k,j} = 1,\ \forall k, j \quad (4\text{-}3)$$

Obviously, the above objective function can be partitioned as K individual problems, and we can optimize each pair $\{\boldsymbol{D}_k, \boldsymbol{\Lambda}_k\}$ separately as

$$J_{\{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k\}}^{(k)} = \arg\min_{\boldsymbol{D}_k,\boldsymbol{\Lambda}_k} \left(\left\| \boldsymbol{P}\boldsymbol{A}_k - \boldsymbol{D}_k\boldsymbol{\Lambda}_k \right\|_2^2 + \lambda_1 \left\| \boldsymbol{\Lambda}_k \right\|_2^2 + \lambda_2 \left\| \boldsymbol{\Lambda}_k - \boldsymbol{\Gamma}_k \right\|_2^2 \right) \quad \text{s.t. } \boldsymbol{d}_{k,j}^T \boldsymbol{d}_{k,j} = 1,\ \forall j \quad (4\text{-}4)$$

$\boldsymbol{D}_k$ and $\boldsymbol{\Lambda}_k$ are also solved alternatively and iteratively. To make the optimization easier, we initialize $\boldsymbol{\Gamma}_k$ as zero, and in the following iterations $\boldsymbol{\Gamma}_k$ can be calculated as the column mean matrix of the updated coefficient matrix $\boldsymbol{\Lambda}_k$. Therefore, $\boldsymbol{\Gamma}_k$ can be viewed as a known constant matrix in optimizing $\boldsymbol{D}_k$ and $\boldsymbol{\Lambda}_k$ in each iteration.

From some initialization of $\boldsymbol{D}_k$ (for example, random initialization), the coding coefficients $\boldsymbol{\Lambda}_k$ can be computed. In one iteration, once $\boldsymbol{D}_k$ is given, we can readily have an analytical solution of $\boldsymbol{\Lambda}_k$ as follows:

$$\boldsymbol{\Lambda}_k = (\boldsymbol{D}_k^T \boldsymbol{D}_k + (\lambda_1 + \lambda_2)\boldsymbol{I})^{-1} (\boldsymbol{D}_k^T \boldsymbol{P}\boldsymbol{A}_k + \lambda_2 \boldsymbol{\Gamma}_k) \quad (4\text{-}5)$$

When $\boldsymbol{\Lambda}_k$ is obtained, the dictionary $\boldsymbol{D}_k$ can then be updated. The procedures of updating $\boldsymbol{D}_k$ are the same as those in [59].

After several iterations, all the $\boldsymbol{D}_k$ and $\boldsymbol{\Lambda}_k$ can be obtained, and we can consequently obtain the whole dictionary $\boldsymbol{D}$ and the associated coefficient matrix $\boldsymbol{\Lambda}$.

Step 3) Fix $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$, update $\boldsymbol{P}$. Let $\boldsymbol{X}=\boldsymbol{D}\boldsymbol{\Lambda}$, the objective function in Eq. (4-2) is reduced to:

$$J_{\boldsymbol{P}} = \arg\min_{\boldsymbol{P}} \left\{ \left\| \boldsymbol{P}\boldsymbol{A} - \boldsymbol{X} \right\|_2^2 - \gamma_1 \left\| \boldsymbol{P}\boldsymbol{A}_t \right\|_2^2 - \gamma_2 \left\| \boldsymbol{P}\boldsymbol{A}_b \right\|_2^2 \right\} \text{ s.t. } \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I} \quad (4\text{-}6)$$

The above sub-objective function $J_{\boldsymbol{P}}$ is itself non-convex, and we can have a local minimum of it as follows. First, since $\boldsymbol{P}\boldsymbol{P}^T=\boldsymbol{I}$, we have

$$\left\|\boldsymbol{PA}-\boldsymbol{X}\right\|_2^2 = tr(\boldsymbol{P}\varphi(\boldsymbol{P})\boldsymbol{P}^T) \tag{4-7}$$

where $\phi(\boldsymbol{P}) = (\boldsymbol{X}-\boldsymbol{P}^T\boldsymbol{A})(\boldsymbol{X}-\boldsymbol{P}^T\boldsymbol{A})^T$. Let $\boldsymbol{S}_t = \boldsymbol{A}_t\boldsymbol{A}_t^T$ and $\boldsymbol{S}_b = \boldsymbol{A}_b\boldsymbol{A}_b^T$, we have $\left\|\boldsymbol{PA}_t\right\|_2^2 = tr(\boldsymbol{PS}_t\boldsymbol{P}^T)$ and $\left\|\boldsymbol{PA}_b\right\|_2^2 = tr(\boldsymbol{PS}_b\boldsymbol{P}^T)$. $J_{\boldsymbol{P}}$ can then be rewritten as

$$\begin{aligned} J_{\boldsymbol{P}} &= \arg\min_{\boldsymbol{P}} \left\{ tr\left(\boldsymbol{P}\varphi(\boldsymbol{P})\boldsymbol{P}^T\right) - \gamma_1 tr\left(\boldsymbol{PS}_t\boldsymbol{P}^T\right) - \gamma_2 tr\left(\boldsymbol{PS}_b\boldsymbol{P}^T\right)\right\} \text{ s.t. } \boldsymbol{PP}^T=\boldsymbol{I} \\ &= \arg\min_{\boldsymbol{P}} \{ tr\left(\boldsymbol{P}\left(\varphi(\boldsymbol{P}) - \gamma_1\boldsymbol{S}_t - \gamma_2\boldsymbol{S}_b\right)\boldsymbol{P}^T\right)\} \end{aligned} \tag{4-8}$$

To solve the above minimization in the current iteration *h*, we use $\varphi(\boldsymbol{P}_{(h-1)})$ to approximate the $\varphi(\boldsymbol{P})$ in Eq. (4-8), where $\boldsymbol{P}_{(h-1)}$ is the projection matrix obtained in iteration *h*-1. Then by using the SVD technique, we have

$$[\boldsymbol{U},\boldsymbol{\Sigma},\boldsymbol{V}] = SVD\left(\varphi(\boldsymbol{P}_{(h-1)}) - \gamma_1\boldsymbol{S}_t - \gamma_2\boldsymbol{S}_b\right) \tag{4-9}$$

where $\boldsymbol{\Sigma}$ is diagonal matrix formed by the eigenvalues of $\left(\varphi(\boldsymbol{P}_{(h-1)}) - \gamma_1\boldsymbol{S}_t - \gamma_2\boldsymbol{S}_b\right)$. Then we can take the updated $\boldsymbol{P}$ as the first *l* most important eigenvectors in $\boldsymbol{U}$, i.e., let $\boldsymbol{P}_{(h)}=\boldsymbol{U}(1{:}l, :)$. However, in this way the update of $\boldsymbol{P}$ may be too big, and make the optimization of the whole system in Eq. (4-2) unstable. Therefore, we choose to update $\boldsymbol{P}$ gradually in each iteration and let

$$\boldsymbol{P}_{(h)}=\boldsymbol{P}_{(h-1)}+c(\boldsymbol{U}(1{:}l, :)-\boldsymbol{P}_{(h-1)}) \tag{4-10}$$

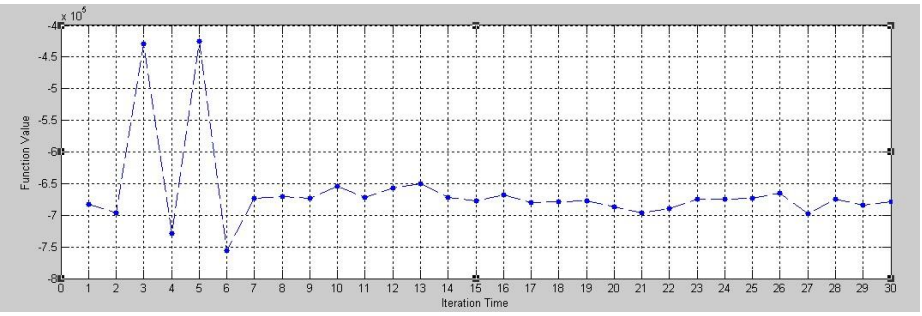
where *c* is a small positive constant to control the change of $\boldsymbol{P}$ in iterations.

Step 4) Stopping criterion. If the maximum iteration number is reached, or the difference between the objective function $J_{\boldsymbol{P},\{\boldsymbol{D}_k,\boldsymbol{A}_k\}}$ in adjacent iterations is smaller a preset value ε, then stop and output $\boldsymbol{P}$ and $\boldsymbol{D}$. Otherwise go back to Step 2.

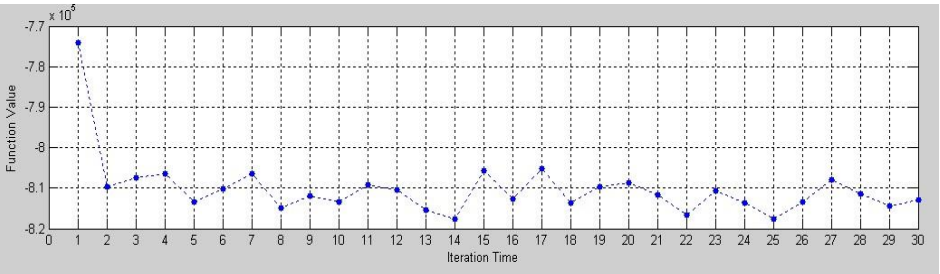
4.3.3 Converge of the JDDRDL Model

The proposed JDDRDL model in Eq. (4-2) is non-convex, and thus the proposed optimization algorithm in Section 4.3.2 can only reach a local minimum of it. Since the

two sub-problems in Step 2 and Step 3 are also non-convex, there is no theoretical guarantee about the convergence of the proposed optimization algorithm in Section 4.3.2. Fortunately, our experiments show that the optimization will lead to a stable solution, though it still has some small oscillation.



(a) AR database



(b) MPIE database

Figure 4.1: The convergence curves of JDDRDL model on the (a) AR and (b) MPIE databases (X axis: Iteration time, Y axis: Function value).

Let's use the AR database [72] and MPIE database [73] as examples to illustrate the optimization process of JDDRDL. The dimensionality of the face images is reduced to 300. The curves of the objective function $J_{\boldsymbol{P},\{\boldsymbol{D}_k,\boldsymbol{A}_k\}}$ vs. the iteration number are plotted in Fig. 4.1(a) and Fig. 4.2(b), respectively, for the two databases. We can see that after several iterations (e.g., 6 iterations), the value of the objective function becomes stable, and it varies only in a small range. Usually, the iteration will stop within 15 times. Our experimental results also show that stopping the minimization with more or less iterations, the resulted projection $\boldsymbol{P}$ and dictionary $\boldsymbol{D}$ will lead to almost the same FR rates. This indicates that although the proposed JDDRDL algorithm cannot lead to an ideal convergence, it is not sensitive to the iteration number. In our experiments, we set the maximal iteration number as 15 and it works well. Here we use PCA to initialize the

projection matrix $\boldsymbol{P}$, however, experiment results show that the model also can converge if we use random projection matrix to initialize $\boldsymbol{P}$, the only difference is that it takes more iterations to converge than initialized by PCA, but the final recognition accuracy does not change. In the following experiments, we adopt PCA to initialize $\boldsymbol{P}$ in order to reduce computational time.

4.3.4 The Classification Scheme

After we obtain the projection matrix $\boldsymbol{P}$, the query sample $\boldsymbol{y}$ can be projected into the lower dimensional space by $\boldsymbol{Py}$, and then the lower dimensional feature $\boldsymbol{Py}$ can be coded over the dictionary $\boldsymbol{D}$. Here we adapted the collaborative representation model with l_2-norm regularization [35] for coding because of its effectiveness and efficiency:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \left\| \boldsymbol{Py} - \boldsymbol{D\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_2^2 \right\} \tag{4-11}$$

where λ is a positive scalar. Obviously, we have $\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{D}^T \boldsymbol{D} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{D}^T \boldsymbol{P} \boldsymbol{y}_0$. The resulted coding vector can be written as $\hat{\boldsymbol{\alpha}} = [\hat{\boldsymbol{\alpha}}_1; \ldots; \hat{\boldsymbol{\alpha}}_k; \ldots; \hat{\boldsymbol{\alpha}}_n]$, where $\hat{\boldsymbol{\alpha}}_k$ is the sub-coding-vector associated with each sub-dictionary $\boldsymbol{D}_i$.

Once the coding vector $\hat{\boldsymbol{\alpha}}$ is computed, the classification can be conducted based on the reconstruction residual of each class, as that in SRC [15] or CRC [35]. However, in the proposed JDDRDL algorithm, the mean of the coding vectors $\boldsymbol{\Lambda}_k$ of each class, denoted by $\boldsymbol{u}_k$, is also learnt, and the distance between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{u}_k$ is also useful for classification, as shown in [84]. Therefore, we adopted the classifier in [84] for the final classification. Let

$$e_k = \left\| \boldsymbol{Py} - \boldsymbol{D}_k \hat{\boldsymbol{\alpha}}_k \right\|_2^2 + \omega \left\| \hat{\boldsymbol{\alpha}}_k - \boldsymbol{\mu}_k \right\|_2^2 \tag{4-12}$$

where ω is the constant to balance the contribution of the two terms. The final classification is performed by identity ($\boldsymbol{y}$) = $\text{argmin}_k\{\boldsymbol{e}_k\}$.

4.4 Experiment Verification

In this section, we use several benchmark face recognition databases to verify the performance of our proposed JDDRDL scheme. The representative algorithms that employ dictionary learning and/or dimensionality reduction under the SRC framework, including SRC [15] with PCA and LDA, CRC [35] with PCA and LDA, metaface learning for SRC (MFL-SRC) [59], dimension reduction for SRC (DR-SRC) [57] and the recently proposed FDDL [84], are used for comparison. The l_1_ls [76] toolbox, which is a stable l_1-minimization solver, is used to solve the l_1-minimization problem (for other l_1-minimization solvers, please see [78]) in the SRC related algorithms. On each database, we first test the robustness of these competing methods to the number of training samples, and then show their results with different dimensionalities of the features. The computation complex is very high to solve $\boldsymbol{P}$ and $\boldsymbol{D}$ in the JDDRDL model, however, $\boldsymbol{P}$ and $\boldsymbol{D}$ can be pre-calculated off-line. When a new test sample $\boldsymbol{y}$ comes, we can project it into a low dimension space and code over $\boldsymbol{D}$ (using the l_2 norm constraint), which is the same as CRC model, so the computational time of JDDRDL is similar with CRC model.

4.4.1 Parameters Selection

There are four parameters (λ_1, λ_2, γ_1 and γ_2) in our JDDRDL model. We set $\lambda_1 = \lambda_2 = 0.005$ by experience in our experiments. The other two parameters, γ_1 and γ_2, will affect the updating of the projection matrix $\boldsymbol{P}$ (see Eq. (4-8)). Our criterion to select them is to keep the three items in Eq. (4-8) be of the similar magnitude. In our experiments, we choose $\gamma_1=10$ and $\gamma_2=1$ in all databases, which could lead to satisfying results. In testing, the scalar λ (refer to Eq. (4-11)) is set as 0.001 and ω (refer to Eq. (4-12)) is set as 0.01 in all experiments by experience.

4.4.2 Face Recognition without Disguise and Occlusion

a) AR database: The AR database [72] consists of over 4,000 frontal images from 126 individuals. For each individual, 26 pictures were taken in two separate sessions. In our

experiments, a subset that contains 50 males and 50 females with 6 illumination and 8 expression variations in two sessions is used (please refer to Fig. 4.2 for some examples).

We randomly chose 2~7 samples per subject for training, while the other samples were used as query samples, all the samples were projected into a 550 dimensional subspace (Samples in LDA+SRC and LDA+CRC schemes were projected into a 99 dimensional subspace). The experiments were repeated 50 times to calculate the average recognition rate and the corresponding variance. The FR rates by competing methods are listed in Table 4.1. It can be seen that when the number of training samples per class is not very small, e.g., 7 samples per class, the recognition rates by all competing methods are quite similar and satisfying. With the decrease of the number of training samples, the recognition rates of all methods drop, especially for LDA+SRC and LDA+CRC. This is mainly because LDA is sensitive to the number of training samples. The proposed JDDRDL achieves the highest FR rates among all the competing methods. Particularly, it is less sensitive to the small sample size problem. When the number of training samples per class is relatively high such as 6 or 7 samples per class, JDDRDL has very close recognition rates to FDDL. However, when the number of training samples is relatively small such as 2~5 samples per class, the difference between the recognition rates of JDDRDL and other methods is getting higher. Overall, JDDRDL's performance is very stable.

Figure 4.2: Some samples from the AR database.

Table 4.1: Recognition rates on the AR database with different number of training samples.

No. of training samples	2	3	4	5	6	7
JDDRDL	0.734±0.037	0.759±0.026	0.818±0.020	0.897±0.017	0.929±0.020	0.941±0.022
DR-SRC	0.711±0.034	0.740±0.028	0.798±0.022	0.871±0.020	0.908±0.021	0.930±0.025
MFL-SRC	0.714±0.031	0.736±0.023	0.790±0.018	0.872±0.024	0.909±0.027	0.932±0.019
PCA+SRC	0.705±0.029	0.731±0.024	0.794±0.014	0.872±0.018	0.910±0.020	0.932±0.018
LDA+SRC	0.494±0.044	0.534±0.033	0.718±0.020	0.859±0.014	0.892±0.027	0.914±0.024
PCA+CRC	0.708±0.030	0.737±0.028	0.788±0.019	0.874±0.021	0.910±0.018	0.930±0.020
LDA+CRC	0.491±0.029	0.534±0.031	0.714±0.028	0.859±0.019	0.890±0.022	0.912±0.015
FDDL	0.690±0.032	0.702±0.029	0.796±0.015	0.888±0.020	0.924±0.022	0.933±0.028

Table 4.2: Recognition rates on the AR database under different feature dimensions.

Dimension	99	350	400	450	500	550
JDDRDL	---	0.805±0.018	0.813±0.023	0.823±0.021	0.822±0.027	0.818±0.020
DR-SRC	---	0.787±0.022	0.791±0.020	0.801±0.024	0.804±0.031	0.798±0.022
MFL-SRC	---	0.788±0.020	0.789±0.014	0.809±0.018	0.798±0.021	0.790±0.018
PCA+SRC	---	0.782±0.027	0.783±0.014	0.804±0.017	0.800±0.025	0.794±0.014
LDA+SRC	0.718±0.020	---	---	---	---	---
PCA+CRC	---	0.784±0.027	0.787±0.020	0.800±0.020	0.793±0.036	0.788±0.019
LDA+CRC	0.714±0.028	---	---	---	---	---
FDDL	---	0.782±0.023	0.794±0.019	0.802±0.024	0.801±0.034	0.796±0.015

We then evaluate the performance of JDDRDL on different dimensionalities. Four samples of each subject are randomly chosen for training, and all the remaining images are used as query images. The recognition rates with different feature dimensions by the competing methods are shown in Table 4.2. JDDRDL surpasses other competing schemes on average. It can be seen that when the dimensionality is relatively low, e.g., 350, all the methods (except for LDA+SRC and LDA+CRC) have similar results. With the increase of feature dimension, e.g., above 450, the proposed JDDRDL shows visible improvement over the other methods.

b) Multi PIE database: The CMU Multi-PIE database [73] contains image of 337 subjects captured in four sessions with simultaneous variations in pose, expression, and illumination. Among these 337 subjects, all the 249 subjects in Session 1 are used (see Fig. 3 for example samples). We randomly selected 2 to 7 samples per subject as our training

set while the other images were used as query set, and projected into a subspace of 550 dimensions (Samples in LDA+SRC and LDA+CRC schemes are projected into a subspace of 248 dimensions). Also, all experiments were repeated for 50 times to calculate the mean and variance of the FR rates. Table 4.3 shows the results by different methods. We can draw similar conclusions to those on the AR database, i.e., the proposed JDDRDL achieves the best FR rates and its advantage over the other methods is more remarkable when the number of training samples is less sufficient.

Table 4.4 lists the recognition rates of the competing methods with different dimensions of features. Four images were randomly chosen from each subject for training set, and the remaining samples were used as for testing, and such experiments were repeated 50 times as well. Similar to what we observed on the AR database, JDDRDL achieves more remarkable improvement over the other methods with the increase of dimensionality. It is also noticed that LDA+SRC and LDA+CRC have good performance on MPIE since MPIE is a large scale dataset with 249 classes, which allows LDA to use enough number of projections to classify the query samples.

Figure 4.3: Some samples from the MPIE database.

Table 4.3: Recognition rates on the MPIE database with different number of training samples.

No. of training samples	2	3	4	5	6	7
JDDRDL	0.756±0.044	0.837±0.029	0.900±0.018	0.906±0.020	0.910±0.011	0.912±0.008
DR-SRC	0.744±0.047	0.824±0.033	0.876±0.024	0.888±0.022	0.902±0.025	0.904±0.010
MFL-SRC	0.741±0.034	0.826±0.020	0.871±0.021	0.881±0.013	0.889±0.012	0.907±0.014
PCA+SRC	0.743±0.039	0.822±0.040	0.880±0.029	0.891±0.028	0.894±0.009	0.905±0.016
LDA+SRC	0.421±0.040	0.795±0.026	0.874±0.020	0.884±0.016	0.895±0.014	0.910±0.009
PCA+CRC	0.745±0.037	0.820±0.033	0.875±0.015	0.893±0.030	0.898±0.013	0.907±0.013
LDA+CRC	0.414±0.042	0.801±0.028	0.877±0.026	0.880±0.019	0.900±0.020	0.908±0.019
FDDL	0.659±0.035	0.810±0.041	0.888±0.017	0.904±0.026	0.908±0.016	0.910±0.015

Table 4.4: Recognition rates on the MPIE database under different feature dimensions.

Dimension	248	350	400	450	500	550
JDDRDL	---	0.866±0.016	0.872±0.011	0.878±0.014	0.886±0.016	0.900±0.018
DR-SRC	---	0.858±0.015	0.867±0.017	0.864±0.010	0.875±0.020	0.876±0.024
MFL-SRC	---	0.853±0.011	0.859±0.010	0.865±0.016	0.871±0.017	0.871±0.021
PCA+SRC	---	0.870±0.014	0.874±0.021	0.867±0.021	0.878±0.018	0.880±0.029
LDA+SRC	0.874±0.020	---	---	---	---	---
PCA+CRC	---	0.873±0.013	0.874±0.014	0.870±0.019	0.877±0.019	0.875±0.015
LDA+CRC	0.877±0.026	---	---	---	---	---
FDDL	---	0.866±0.011	0.871±0.012	0.872±0.014	0.881±0.016	0.888±0.017

c) Extended Yale B Database: The extended Yale B [56] database contains about 2,414 frontal face images of 38 individuals taken under varying illumination conditions. We randomly chose 2 to 7 images from each person as training set, and used the rest images as testing set. Similarly, all the samples were projected into a subspace of 550 dimensions (Samples in LDA+SRC and LDA+CRC schemes are projected into a subspace of 37 dimensions) and the experiments were repeated 50 times. The FR results are shown in Table 4.5.

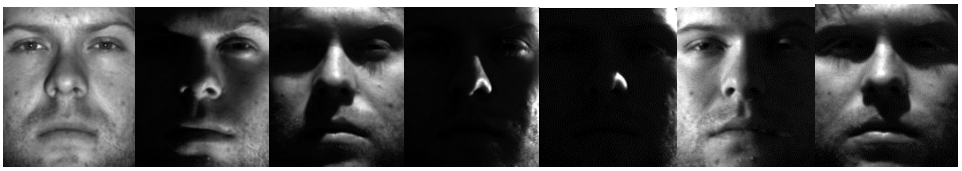


Figure 4.4: Some samples from the Extended Yale B database.

Table 4.5: Recognition rates on the Yale B database with different number of training samples.

No. of training samples	2	3	4	5	6	7
JDDRDL	0.549±0.034	0.653±0.036	0.674±0.025	0.682±0.022	0.696±0.030	0.705±0.024
DR-SRC	0.530±0.038	0.636±0.031	0.656±0.030	0.671±0.025	0.689±0.023	0.698±0.021
MFL-SRC	0.534±0.029	0.631±0.025	0.657±0.026	0.668±0.023	0.690±0.023	0.692±0.018
PCA+SRC	0.535±0.031	0.641±0.034	0.652±0.024	0.670±0.029	0.687±0.024	0.690±0.031
LDA+SRC	0.462±0.032	0.532±0.031	0.603±0.028	0.665±0.030	0.681±0.019	0.681±0.022
PCA+CRC	0.532±0.028	0.644±0.024	0.650±0.022	0.671±0.025	0.685±0.024	0.692±0.025
LDA+CRC	0.460±0.039	0.535±0.033	0.609±0.031	0.662±0.028	0.679±0.020	0.682±0.014
FDDL	0.441±0.042	0.538±0.037	0.636±0.023	0.675±0.021	0.693±0.017	0.701±0.025

Table 4.6: Recognition rates on the Extended Yale B database under different feature dimensions.

Dimension	37	350	400	450	500	550
JDDRDL	---	0.658±0.017	0.660±0.015	0.665±0.021	0.666±0.031	0.674±0.025
DR-SRC	---	0.644±0.019	0.647±0.017	0.648±0.022	0.651±0.028	0.656±0.030
MFL-SRC	---	0.640±0.022	0.640±0.025	0.642±0.029	0.645±0.033	0.657±0.026
PCA+SRC	---	0.640±0.026	0.641±0.019	0.644±0.018	0.650±0.026	0.652±0.024
LDA+SRC	0.603±0.028	---	---	---	---	---
PCA+CRC	---	0.637±0.014	0.645±0.022	0.649±0.023	0.652±0.024	0.650±0.022
LDA+CRC	0.609±0.031	---	---	---	---	---
FDDL	---	0.614±0.019	0.616±0.024	0.618±0.025	0.624±0.028	0.636±0.023

We then randomly selected 4 images from each subject as the training set, and took the remaining samples as the testing set. The FR rates under different dimensions are shown in Table 4.6. Compared with the AR database, the extended Yale B database has less expression variations but larger illumination changes (please see Fig. 4.4 for examples). When the number of training samples is insufficient, the FR becomes very challenging due to the large variation in illumination. From Tables 4.5 and 4.6, one can see that the proposed JDDRDL method achieves the highest recognition rates among the competing schemes. When there are only 3 training samples per subject, JDDRDL achieves about 10% higher recognition rate than FDDL, which is a state-of-the-art discriminative dictionary learning method. This is because FDDL performs dimensionality separately from the discriminative dictionary learning process so that it requires enough training samples to stably compute the statistics. By coupling the dimensionality reduction and dictionary learning processes, the proposed JDDRDL can increase much the robustness to the number of training samples while yielding a discriminative dictionary.

d). FERET database: A pose subset of the FERET database [86] is used here, which includes the frontal face images marked with "*ba*", "*bj*", and "*bk*". Since there are only three samples for each subject, in each experiment we use two samples for training and the other one for testing. In the first experiment, the image marked with "*ba*" and "*bj*" were used as training samples, and totally 200 classes and 400 samples were used in the training set. The testing set includes the images marked with "*bk*" for each subject (please refer to

Fig. 4.5 for examples). The FR result is shown in Table 4.7(a). In experiment 2, the images marked with "*ba*" and "*bk*" were used for training and "*bj*" was used for testing. The result is list in Table 4.7(b). Similarly, in the third experiment, images "*bj*" and "*bk*" were used for training and "*ba*" was used for testing. The result is list in the Table 4.7(c).

Figure 4.5: Some samples from the FERET database.

Table 4.7(a): Recognition rates on the FERET database under different feature dimensions.

Dimension	199	350	400	450	500	550
JDDRDL	---	0.795	0.810	0.800	0.785	0.780
DR-SRC	---	0.790	0.790	0.785	0.785	0.775
MFL-SRC	---	0.795	0.795	0.785	0.770	0.770
PCA+SRC	---	0.785	0.785	0.790	0.785	0.775
LDA+SRC	0.715	---	---	---	---	---
PCA+CRC	---	0.790	0.795	0.790	0.785	0.775
LDA+CRC	0.715	---	---	---	---	---
FDDL	---	0.720	0.725	0.725	0.715	0.715

Table 4.7(b): Recognition rates on the FERET database under different feature dimensions.

Dimension	199	350	400	450	500	550
JDDRDL	---	0.895	0.900	0.900	0.895	0.895
DR-SRC	---	0.880	0.880	0.875	0.875	0.875
MFL-SRC	---	0.875	0.875	0.875	0.880	0880
PCA+SRC	---	0.890	0.890	0.890	0.885	0.880
LDA+SRC	0.730	---	---	---	---	---
PCA+CRC	---	0.890	0.895	0.890	0.890	0.885
LDA+CRC	0.735	---	---	---	---	---
FDDL	---	0.790	0.795	0.795	0.800	0.800

Table 4.7(c): Recognition rates on the FERET database under different feature dimensions.

Dimension	199	350	400	450	500	550
JDDRDL	---	0.915	0.930	0.940	0.920	0.920
DR-SRC	---	0.895	0.905	0.920	0.910	0.910
MFL-SRC	---	0.895	0.900	0.915	0.910	0.905
PCA+SRC	---	0.895	0.900	0.910	0.910	0.905
LDA+SRC	0.755	---	---	---	---	---
PCA+CRC	---	0.900	0.905	0.910	0.905	0.905
LDA+CRC	0.750	---	---	---	---	---
FDDL	---	0.790	0.800	0.815	0.810	0.810

Similar to the results in other databases, from Fig. 4.7(a), 4.7(b) and 4.7(c) we see that the proposed JDDRDL achieves the highest recognition results in the three experiments, which demonstrates its capability to handle the small sample size problem. The LDA+SRC, LDA+CRC and FDDL methods do not work well on this dataset because their sensitivity to the number of training samples. Compared with DR-SRC, MFL-SRC, PCA+CRC and PCA+SRC, the JDDRDL can always achieve certain improvement in the three experiments.

4.4.3 Face Recognition with Real Disguise and Occlusion

In this section, we consider to apply the JDDRDL to face recognition with occlusion. In order to handle the occlusion problem and improve the recognition rate, we adopt the block partition strategy. It is known that the occlusion usually falls on some patches of the face image, so it is intuitive for us to partition the image into blocks and process each block independently. The results for individual blocks are then aggregated by voting. If some blocks have large reconstruction errors, those blocks are supposed to be largely occluded and they will be discarded.

We partition each training sample into l blocks, producing a set of matrices $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$,…, $\boldsymbol{X}^{(l)}$. (Please refer to Fig. 4.6). Similarly, each query image $\boldsymbol{y}$ is also partitioned into l blocks, e.g., $\boldsymbol{y}^{(1)}$, $\boldsymbol{y}^{(2)}$,…, $\boldsymbol{y}^{(l)}$. we rewrite the k^{th} block of the query sample as a linear combination $\boldsymbol{X}^{(k)}\boldsymbol{\alpha}^{(k)}$, e.g., $\boldsymbol{y}^{(2)}= \boldsymbol{X}^{(k)}\boldsymbol{\alpha}^{(k)}+\boldsymbol{e}^{(k)}$, as that in the block based SRC model.

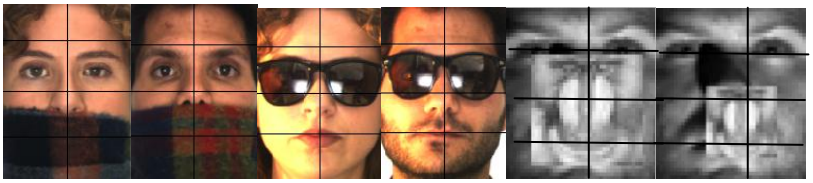


Figure 4.6: Examples of partition samples of real disguise and occluded faces.

Under the JDDRDL, the dictionary is learnt through each block of the training sample while the projection matrix is learnt globally. Suppose we divide each training sample into *l* blocks, the optimization function is:

$$J(\boldsymbol{P},\boldsymbol{D},\boldsymbol{\Lambda}) = \arg\min_{\boldsymbol{P},\boldsymbol{D},\boldsymbol{\Lambda}} \sum_{i=1}^{n}\sum_{k=1}^{l}\{\left\|\boldsymbol{P}\boldsymbol{X}_i^k - \boldsymbol{D}_i^k\boldsymbol{\Lambda}_i^k\right\|_2^2 + \lambda_1\left\|\boldsymbol{\Lambda}_i^k\right\|_2^2 + \lambda_2\left\|\boldsymbol{\Lambda}_i^k - \boldsymbol{\Gamma}_i^k\right\|_2^2 - \gamma_1\left\|\boldsymbol{P}\boldsymbol{A}_t\right\|_2^2 - \gamma_2\left\|\boldsymbol{P}\boldsymbol{A}_B\right\|_2^2\} \quad (4\text{-}13)$$
$$\text{s.t. } \boldsymbol{d}_i^T\boldsymbol{d}_i = 1,\ \forall i,\ \ \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}$$

The meaning of each parameter is similar with the ones in Eq. (4-2) but in a local patch. In our experiment, we divide all samples into 2*4 blocks, and each sub-dictionary is learnt through corresponding blocks. After each sub-dictionary is obtained, the blocks in the query image will be represented via those sub-dictionaries. The following procedures are the same as stated in Section 4.3. Each block is projected into a subspace of 50 dimensions.

The tables below indicate the competing experiment results in the occlusion databases.

Table 4.8: Recognition results by different methods on the AR database with sunglasses and scarves disguise.

Algorithm	Sunglasses	Scarves
SRC [7]	87.0%	59.5%
SRC(partitioned) [7]	97.5%	93.5%
RSC [49]	99.0%	97.0%
JDDRDL	89.5%	61.0%
JDDRDL(partitioned)	98.5%	94.5%

Table 4.9: Recognition results by different methods on the Extended Yale B with various random occlusion ratios.

Occlusion ratio	10%	20%	30%	40%	50%
SRC [7]	100%	99.8%	98.5%	90.3%	65.3%
SRC(partitioned) [7]	100%	99.8%	98.5%	90.7%	66.8%
RSC [49]	100%	100%	99.8%	96.9%	83.9%
JDDRDL	100%	99.8%	99.2%	90.5%	67.1%
JDDRDL(partitioned)	100%	99.8%	99.4%	92.4%	68.7%

Without partition, SRC achieves relatively low recognition result when the disguise or occlusion part covers large part of the face image. In scarves disguise and 50% random block occlusion, the recognition rates of SRC are only 59.5% and 65.3%, respectively, which indicates that the outlier parts seriously affect the reconstruction of the test image. JDDRDL also achieves relatively low recognition rate, but still surpasses SRC. However, partition can greatly increase the discriminative ability for both SRC and JDDRDL. SRC with partition could achieves 97.5% and 93.5% in glasses and scarves disguise, respectively, which are 10.5% and 34% higher than SRC without partition, respectively. With partition, the proposed JDDRDL model also achieves better result than the SRC model, reaching 98.5% and 94.5% on those two datasets, which indicates that the JDDRDL model could also handle the face occlusion problem well and the discriminative information is increased by the joint learning. In random block occlusion dataset, the improvement brought by the partition is not as remarkable as in real disguise. This is mainly because the occlusion ratio is larger, and thus even the test image is partitioned, the portioned patches still contain large occlusion. However, the JDDRDL still surpasses SRC by 1.9% when the occlusion ratio is 50% and 1.7% when the occlusion ration is 40%. RSC achieves the highest recognition rate in all condition, but we want to emphasize that the computation of RSC is very high.

4.5 Conclusion

In this chapter we propose a joint discriminative dimensionality reduction and dictionary learning (JDDRDL) scheme for face recognition. Unlike many methods which focus on dictionary learning (DL) and use PCA or LDA for dimensionality reduction (DR), JDDRDL considers the interaction between DR and DL procedures by coupling them into a unified framework for energy minimization. The DR matrix projects the data into a lower dimensional subspace where the total scatter and between-class scatter of the training data are maximized, while the learnt dictionary associated with the DR matrix is ensured to have a strong representative ability. The JDDRDL incorporates the Fisher Criteria to maximize the between-class scatter and minimize the within-class scatter, JDDRDL does not simply follow the LDA criteria, the learnt dictionary and projection matrix also contribute to the recognition, so the experiment results surpass *+SRC although LDA+SRC is not as good as others. In classification, both the representation residual and the distance between the coding vector and the mean vector of each class are considered. The experimental results on representative face databases demonstrate that the proposed JDDRDL method surpasses many state-of-the-arts face recognition methods.

Chapter 5. Conclusion

5.1 Summary and Contribution of this Thesis

This thesis studies the problem of face recognition with the state-of-art sparse coding and dictionary learning method as the main point of discussion. In fact, many of face recognition schemes deprive from the SRC model. The linear combination through training samples has great reconstruction powers to represent unknown testing samples. Based on this fact, many dictionary learning algorithms are proposed to train a representative dictionary to represent original training and remove noise and other trivial information that can compromise representation and recognition. The main contribution of this work is its proposal of a fast and robust face recognition scheme (i.e., CRMLAM) to deal with face occlusion, and a joint learning method (i.e., JDDRDL) which enables joint learning optimization.

In CRMLAM, the training dataset is represented by a learnt dictionary. The dictionary is used to reconstruct the training dataset to obtain reconstruction error. When the test sample contains occlusion, the distribution of reconstruction error differs from the normal distribution obtained from clear images. With this statistics feature, occlusion positions can be located. The detected occlusion points will be discarded in the recognition procedure to reduce the negative effect of the occlusion parts. Another feature of CRMLAM is that it is based on l_2-norm constraint and thus its computational complexity is quite low. Our experiments show that CRMLAM can be applied to real time systems.

A JDDRDL scheme is also proposed in this thesis. Dimensionality reduction is always a major problem in face recognition. Considering the discriminative ability of dimensionality reduction projector and the dictionary, a joint learning scheme aiming to jointly learn projection matrix and dictionary is studied and verified. The Fisher Criteria motivates us to reduce the within class distance while enlarge the between class distance, and thus the criteria is introduced to the objective function. The learnt projection matrix is able to

impose this feature on newly projected samples, and these samples are also well represented by the dictionary. The experiment shows that this scheme can increase the recognition rate compared with other dimensionality reduction and dictionary learning schemes, especially in terms of small sample size. However, the CRMLAM model is still based on the SRC/CRC framework, which requires sufficient training samples, also the goal of proposing the CRMLAM is to develop a fast yet effective enough occlusion detection method, so we expect the performance of CRMLAM will be quite similar with SRC/CRC when dealing with the small sample size problem and it will not be so remarkable as the performance in the face occlusion problem.

5.2 Future Work

Although the two proposed schemes receive satisfactory results, there is still room for improvement. The CRMLAM scheme aims to solve the common problems of face recognition. Our algorithm can be tested in other databases which are not obtained under controlled laboratory conditions. Another thing is that CRMLAM uses simple criteria to detect the occlusion points with rough estimation of distribution. When reconstruction error is larger than a certain value, the corresponding points will be discarded in the testing sample. In other words, it can be assumed that the weights of these pixels are reset to zero in recognition while the weights of non-discarded pixels remain one. In fact, not all removed pixels are occluded. Some of them are actually “clean” because of wrong judgment. Meanwhile, some pixels, whose reconstruction errors are near the critical value, are mistaken to be non-occluded. These situations should be avoided so that application of a more dedicate weight assignment scheme is taken into consideration instead of simply reserving or removing.

It is observed that the distribution of reconstruction error is nearly Gaussian. According to our assumption, it is highly possible that the points are occluded if reconstruction error is far away from the mean of the Gaussian distribution. As a result, each point is assigned a weight according to its distance from the mean. The range is between zero and one. This is called “soft masking”, which can improve the process of occlusion detection.

Convergence in JDDRDL remains unsolved. Although our experiments show that converge curves do not affect the recognition rates, a more “elegant” solution to our model is still demanded. To better verify its discriminative ability, JDDRDL model can be applied to different tasks, such as digital recognition and gender recognition. We believe JDDRDL can also achieve promising results in those tasks. Our current JDDRDL model is a kind of global projection method. However, as local features are also effective information to describe face images, our objective function can be modified to verify whether we can achieve better result when optimizing locally.

Bibliography

[1] A.K. Jain, P.J. Flynn, and A. Ross, *Handbook of Biometrics*. Springer, 2007.

[2] Biometrics, http://en.wikipedia.org/wiki/Biometrics

[3] BCC Research: http://www.bccresearch.com/report/IFT042B.html

[4] A.K. Jain, R. Bolle, and S. Pankanti, editors, *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, 1999.

[5] K. Delac, M. K. Delac, and M. Grgic, *Face Recognition*. I-Tech Education and Publishing, 2007.

[6] H. Wechsler, *Reliable Face Recognition Methods - System Design, Implementation and Evaluation*. Springer, 2006.

[7] S.Z. Li and A.K. Jain, *Handbook of Face Recognition*. Springer, 2005.

[8] Y. Fang, T. Tan, and Y. Wang, "Fusion of global and local features for face verification", in *Proc. ICPR*, pp. 382-385, 2002.

[9] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminate model for face recognition", *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467-476, 2002.

[10] A. Bronstein, M. Bronstein, and R. Kimmel, "Three-dimensional face recognition", *Int. J. Computer Vision*, vol. 64, no. 1, pp. 5-30, 2005.

[11] A. Bronstein, M. Bronstein, and R. Kimmel, "Expression-invariant representation of faces", *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 188-197, 2007.

[12] S. Gundimada and V.K. Asari, "A novel neighborhood defined feature selection on phase congruency images for recognition of faces with extreme variations", *Int. J. Information Technology*, vol. 3, no. 1, pp. 25-31, 2007.

[13] L. Shen and L. Bai, "A review on Gabor wavelets for face recognition", *Pattern Anal. Appl.*, vol. 9, no. 10, pp. 273-292, 2006.

[14] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition", *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 57-68, 2007.

[15] J. Wright, A.Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210-227, 2009.

[16] Y. Su, S. Shan, X. Chen, and W. Gao, "Hierarchical ensemble of global and local classifiers for face recognition", *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1885-1896, 2009.

[17] S. Xie, S. Shan, X. Chen, and J. Chen, "Fusing local patterns of Gabor magnitude and phase for face recognition", *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1349-1361, 2010.

[18] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition", in *Proc. CVPR*, pp. 84–91, 1994.

[19] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 24, no. 6, pp. 780–788, 2002.

[20] B. Scholkopf, A. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation*, vol. 10, no. 5, pp. 1299-1319, 1998.

[21] K. Kim, K. Jung, and H. Kim, "Face recognition using kernel principal component analysis", *IEEE Signal Processing Letters*, vol. 9, no. 2, pp. 40-42, 2002.

[22] M. Yang, N. Ahuja, and D. Kriegman, "Face recognition using kernel eigenfaces", in *Proc. ICIP*, pp. 37-40, 2000.

[23] C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 26, no. 5, pp. 572-581, 2004.

[24] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711-720, 1997.

[25] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminative analysis of principal components for face recognition", in *Proc. ICAFG*, pp. 336-341, 1998.

[26] W. Zhao, R. Chellappa, and P.J. Phillips, "Subspace linear discriminate analysis for face recognition", Tech. Report, CAR-TR-914, Center for Automation Research, University of Maryland, MD, 1999.

[27] L. Chen, H. Liao, M. Ko, J. Lin, and G. Yu, "A new LDA-based face recognition system which can solve the small sample size problem", *Pattern Recognition*, vol. 33, no. 10, pp. 1713-1726, 2000.

[28] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition", *Pattern Recognition*, vol. 34, no. 10, pp. 2067-2070, 2001.

[29] W. Zuo, D. Zhang, J. Yang, and K. Wang, "BDPCA plus LDA: a novel fast feature extraction technique for face recognition", *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 4, pp. 946-953, 2006.

[30] J. Lu, K. Plataniotis, and A. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms", *IEEE Trans. Neural Network*, vol. 14, no. 1, pp. 117-126, 2003.

[31] X. Liu, W. Chen, P. Yuen, and G. Feng, "Learning kernel in kernel-based LDA for face Recognition under illumination variations", *IEEE Signal Processing Letters*, vol. 16, no. 12, pp.

1019-1022, 2009.

[32] G. Baudat and F. Anouar "Generalized discriminate analysis using a kernel approach", *Neural Computing*, vol. 12, no. 10, pp. 2385-2404, 2000.

[33] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary", in *Proc. ECCV*, pp. 448-461, 2010.

[34] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition", *Proceedings of IEEE*, vol. 98, no. 6, pp. 1031-1044, 2010.

[35] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?", in *Proc. ICCV*, pp. 471-478 2011.

[36] R. Brunelli and T. Poggio, "Face recognition: features versus templates", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 15 no. 10, pp. 1042–1052, 1993.

[37] I. Cox, J. Ghosn, and P. Yianilos, "Feature-based face recognition using mixture-distance", in *Proc. CVPR*, pp. 209-216, 1996.

[38] E. Hjelmas, "Feature-based face recognition", in *Proc. Norwegian Image Processing and Pattern Recognition Conference*, 2000.

[39] H. Eknel, and R. Stiefelhagen, "Local appearance based face recognition using discrete cosine transform", in *Proc. EUSIPCO*, 2005.

[40] T.Ahonen, A.Hadid, and M. Pietik, "Face description with local binary patterns: application to face recognition", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037-2041, 2006.

[41] A. Lantitis, C. Taylor, and T. Cootes, "Automatic face identification system using flexible appearance models", *Image and Vision Computing*, vol. 13, no. 5, pp. 393-401, 1995.

[42] R. Huang, V. Pavlovic, and D. Metaxas, "A hybrid face recognition method using Markov random fields", in *Proc. ICPR*, pp. 157-160, 2005.

[43] S. Lawrence, C. Giles, A. Tsoi, and A. Back, "Face recognition: A hybrid neural network approach", *Technical Report*, UM Computer Science Department, 1998.

[44] D. Gorodnichy, W. Armstrong, and X. Li, "Adaptive logic networks for facial feature detection", *Image Analysis and Processing*, vol. 1311, pp. 332-339, 1997.

[45] S. Kang, K. Young, and R. Park, "Hybird approaches to front view face recognition using the hidden markov model and neural network", *Pattern Recognition*, vol. 31, no. 3, pp. 283-293, 1998.

[46] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang, "Learning with l_1-graph for image analysis", *IEEE Trans. Image Process.*, vol. 19, no.4, pp. 858-866, 2010.

[47] S. Gao, I. Tsang, and L. Chia, "Kernel sparse representation for image classification and face recognition", in *Proc. ECCV*, pp. 1-14, 2010.

[48] L. Zhang, W. Zhou, P. Chang, J. Liu, Z. Yan, T. Wang, and F. Li, "Kernel sparse representation-based classifier", *IEEE Trans. Signal Processing*, vol. PP, vol. 99, pp. 1-1, 2011.

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

[50] Tony Ao Ieong Wai Heng, "iJADE face recognition – A multi-agent based pose and scale invariant human face recognition system", Master thesis, the Hong Kong Polytechnic University, 2006.

[51] A.K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4-20, 2004.

[52] J. Fagertun, "Face Recognition", Master's thesis, Technical University of Denmark, 2005.

[53] Y. Xue, "Non-negative matrix factorization for face recognition", PhD thesis, Hong Kong Baptist University, 2007.

[54] W. Zhao and R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey", *ACM Comput. Surv.*, vol. 35, iss. 4, pp. 399–458, 2003.

[55] M. Turk and A.P. Pentland, "Face recognition using eigenfaces", in *Proc. CVPR*, pp. 302-306, 1991.

[56] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, 2001.

[57] L. Zhang, M. Yang, Z. Feng, and D. Zhang, "On the dimensionality reduction for sparse representation based face recognition", in *Proc. ICPR*, pp. 1237-1240, *2010.*

[58] P. Viola and M. J. Jones, "Robust face recognition via accurate face alignment and sparse representation", in *Proc. International Conference on Digital Image Computing: Techniques and Applications*, pp. 262-269, 2010.

[59] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface Learning for sparse representation based face recognition", in *Proc. ICIP*, pp. 1601-1604, 2010.

[60] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intra-class variant dictionary*", IEEE Trans. Patt. Anal. Mach. Intell.*, preprint, 2012.

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

[62] P. Buyssens and M. Revenu, "Learning sparse face features: Application to face verification", in *Proc. ICPR*, pp. 670-673, 2010.

[63] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images", in *Proc. CVPR*, pp. 763-770, 2010.

[64] A. Yang, Z. Zhou, Y. Ma, Sastry, and S. Shankar, "Towards a robust face recognition system using compressive sensing", in *Proc. INTERSPEECH*, pp. 2250-2253, 2010.

[65] K. Daniilidis, P. Maragos, and N. Paragios, "Kernel sparse representation for image classification and face recognition", in *Proc. ECCV 2010*, pp. 1–14, 2010.

[66] J.P. Brunet, P. Tamayo, T.R. Golun, and J. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization", *Proceedings of the National Academy of Sciences* USA, vol.101, no.12, pp. 4164-4169, 2004.

[67] J. Yang, J. Wang, and T. Huang, "Learning the sparse representation for classification", in *Proc. ICME*, pp. 1-6, 2011.

[68] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images", *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[69] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis", in *Proc. CVPR*, pp. 1-8, 2008.

[70] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin, "Non-parametric Bayesian dictionary learning for sparse image representations", in *Proc. NIPS*, 2009.

[71] D. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition", in *Proc. CVPR*, pp 1-8, 2008.

[72] A.M. Martinez and R. Benavente, "The AR Face Database", *CVC Technical Report* No. 24, 1998.

[73] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE", *Image and Vision Computing*, vol. 28, pp. 807–813, 2010.

[74] 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.

[75] Hestenes, Magnus R. Stiefel, and Eduard, "Methods of Conjugate Gradients for Solving Linear Systems", *Journal of Research of the National Bureau of Standards*, vol. 49, no. 6, 1952.

[76] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares", *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[77] D. Malioutove, M. Cetin, and A. Willsky, "Homotopy continuation for sparse signal representation", in *Proc. ICASSP*, pp. 733-736, 2005.

[78] A. Y. Yang, A. Ganesh, Z. H. Zhou, S. S. Sastry, and Y. Ma, "Fast l_1-minimization algorithms and application in robust face recognition", UC Berkeley, *Technique Report*, 2010.

[79] I. Tosic and P. Frossard, "Dictionary learning", *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27-38, 2011.

[80] M. Aharon, M. Elad, and A. M. Bruckstein, "The K-SVD: An algorithm for designing of over-

complete dictionaries for sparse representation", *IEEE Trans. Signal Processing*, vol. 51, no. 11, pp. 4311-4322, 2006.

[81] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning", in *Proc. NIPS*, 2008.

[82] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries", *Technical report*, University of Minnesota, 2007.

[83] R. Rubinstein, A.M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling", *Proceedings of IEEE, Special Issue on Applications of Compressive Sensing & Sparse Representation*, vol. 98, no. 6, pp. 1045-1057, 2010.

[84] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation", in *Proc. ICCV*, pp. 543-550, 2011.

[85] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms", *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090-1104, 2000.

[86] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition", *in Proc. CVPR*, pp. 2691-2698, 2010.

[87] O. Bryt and M. Elad, "Compression of facial images using the K-SVD algorithm", *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 270-282, 2008.