FACIAL IMAGE ANALYSIS AND RECOGNITION IN THE WILD

MUHAMMAD SAAD SHAKEEL

PhD

The Hong Kong Polytechnic University

2019

THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF ELECTRONIC AND INFORMATION ENGINEERING

Facial Image Analysis and Recognition in the Wild

Muhammad Saad Shakeel

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

August 2018

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 or diploma, except where due acknowledgment has been made in the text.

_____ (Signed)

Muhammad Saad Shakeel (Name of Student)

Abstract

The human face is the most widely used biometric for recognition and verification of one's identity. It has been widely studied and analyzed in the past few decades due to its various advantages over other biometrics. In the past few years, face recognition research has reached many milestones, due to the availability of large amounts of training data and high computational power. Researchers have already achieved more than 99% recognition accuracy on one of the most challenging face datasets, namely, Labeled Faces in the Wild (LFW). In spite of this, some challenges still remain in the areas of low-resolution (LR), and age-invariant face recognition. Moreover, none of the research works have investigated the problem of noise variations in cross-age face images. The major objective of this thesis is to develop efficient algorithms that can handle and overcome these major challenges.

In this thesis, we have first proposed a sparse-coding based method, which aims to recognize low-resolution face images up to the size of 8×8 captured under controlled and uncontrolled environments. We first down-sample gallery faces to the same resolution as a query image, and then extract effective local features, namely Gabor wavelets, and local binary pattern difference feature. Extracted features are then decomposed into a low-rank feature matrix, and a sparse error matrix. After that, a sparse coding-based objective function is proposed that projects learned gallery and query face images onto a discriminant low-dimensional sparse feature subspace for recognition. Our method preserves the structural information while projecting samples onto a new feature subspace, which results in the accurate classification. Our method provides state-of-the-art performance in recognizing very LR images, and outperforms both conventional and deep learning-based face recognition methods.

In the second part of this thesis, we investigate the existing work for solving age-invariant face recognition problem. A typical approach to solving the aging problem is to synthesize a test image to be the same age as a gallery image, and then perform recognition. However, development of an accurate aging model requires strong parametric assumptions and also a large amount of training data, which makes it unsuitable for real-world applications. Another approach, based on discriminative models, aims to learn high-level facial features invariant to age progression. In this thesis, we have proposed a robust deep-feature encoding-based discriminative model for aging face recognition. First, deep features are learned using a pre-trained deep convolutional neural network (AlexNet), which are then encoded using our proposed locality-constraint feature-encoding framework. By incorporating the locality information, correlation between the features of the same identity can be well captured by sharing the local bases of the learned

codebook. To make the codebook discriminative in terms of age-progression, canonical correlation analysis (CCA) is utilized to fuse the pair of training set features with large age gaps. Encoded features are then passed to the linear-regression based classifier for recognition. Our proposed method does not require any age-label information for recognition purposes.

Aging variation is a complex non-linear process, which affects various facial regions over a period of time. However, the periocular region of a human face contains complex biomedical features, such as eyebrows, contour, eyeballs, eyelids, etc. that vary very little with time. Furthermore, the available training and testing data might be corrupted with some random noise. Previous methods assume that training data is collected under controlled environments, which then degrade their performance, when corrupted testing data is presented for recognition. To solve this problem, we have proposed a manifold-constrained low-rank decomposition algorithm, which recovers underlying identity information from corrupted data samples to provide better feature representation. Furthermore, our method also preserves the local structure of the data samples, while removing the sparse errors. The resultant low-rank feature matrix is then encoded by learning an age-discriminative codebook using our proposed feature encoding-based framework. Since CCA cannot model the non-linear relationship between the two data samples, we utilize kernel canonical correlation analysis (KCCA) to fuse the pair of training set's features with large age differences, which are then used to learn an age-discriminative codebook. Encoded features are then passed to the nearest neighbor classifier for recognition. Performance of our proposed method is evaluated using both the whole face region and the periocular region with different levels of corrupted pixels in both training and testing data. Our proposed method proves to be highly robust against different levels of noise variations, and provides superior performance in terms of recognition rate.

All the proposed methods in this thesis are evaluated by conducting extensive sets of experiments on challenging face datasets. Furthermore, proposed methods are also compared with other state-of-the-art face-recognition methods.

List of Publications

[1] M. Saad Shakeel, and Kin-Man Lam, "Recognition of Low-Resolution face images using sparse coding of Local Features," Proceedings, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA'2016)*, December 2016, Jeju, South Korea, pp. 1-6.

[2] M. Saad Shakeel, Kin-Man Lam, and Shun-Cheung Lai, "Learning Sparse Discriminant Low-Rank Features for Low-Resolution Face Recognition," *revised manuscript submitted to Journal of visual communication and Image representation*, January 2019.

[3] M. Saad Shakeel, and Kin-Man Lam, "Deep Feature encoding-based discriminative model for Age-invariant face recognition*," revised manuscript prepared to submit to Pattern Recognition*, 2019.

[4] M. Saad Shakeel, and Kin-Man Lam, "Deep low-rank feature learning and encoding for cross-age face recognition," *revised manuscript being prepared to submit to IEEE transactions on information forensics and security*, 2019.

Acknowledgment

Thanks to Almighty ALLAH for providing me the courage, determination, and the countless blessings that helps me in finishing my research work.

I would like to express my deepest thanks to my supervisor Prof. Kin-Man Lam for his continuous support and guidance throughout my PHD studies. He not only enlightens me with his research knowledge and experience, but also motivates me during the tough time.

I would like to dedicate my work to my beloved parents, wife, and sisters, who gave me the strength and courage for what I achieved in my life. Their unconditional love was the biggest support throughout my PHD studies.

I am highly grateful to all my colleagues with whom I worked in the past three years: Huiling Zhou, Hailiang Li, Cigdem Turan, Shun-Cheung Lai, and Chenhang He. I am thankful to all of them especially Shun-Cheung Lai, who shares some useful concepts regarding my field of research.

Table of Contents

List of Abbreviations

AIFR Age-invariant face recognition

ANN Artificial neural network

CAFR Cross-age face recognition

CCA Canonical correlation analysis

CNN Convolutional neural network

DCT Discrete cosine transform

DSIFT Densely-sampled scale invariant feature transform

EM Expectation-maximization

FDA Fisher discriminant analysis

FRS Face recognition system

GOP Gradient orientation pyramid

GW Gabor wavelets

HOG Histogram of oriented gradients

HR High resolution

ICA Independent component analysis

ISOMAP Isometric mapping

KCCA Kernel canonical correlation analysis

LBPD Local binary pattern difference

LDA Linear discriminant analysis

LLC Locality-constrained linear coding

LLE Locally linear embedding

LR Linear-regression

MDS Multi-dimensional scaling

MLBP Multi-scale local binary pattern

MR Medium resolution

NN Nearest neighbor

NPE Neighborhood preserving embedding

PCA Principal Component analysis

RBF Radial basis function

SRC Sparse representation-based classifier

SR Super-resolution

SVM Support Vector Machine

List of Figures

List of Tables

Chapter 1 Introduction

The human face is a widely used biometric for solving many identification and verification problems. It has attracted considerable amount of attention from the researchers around the world due to its useful applications. The main objective of this chapter is to give a brief introduction to facial image analysis and its real-world applications. Moreover, the motivation behind our research work, and existing challenges in face-recognition research, will also be discussed. Furthermore, we also discuss the methodologies proposed in our thesis, along with a brief outline and discussion.

1.1 Research Background

The human face, which is considered the most vital part of the human body, contains a lot of useful information, regarding behavior, identity, expressions, age, etc. It has many useful applications, such as criminal identification, finding missing children, security monitoring, etc. Although human beings have excellent capabilities for recognizing an unknown person from his/her face, now due to the evolution of artificial intelligence, it is possible for computers to meet human-level performance in many computer-vision applications. Unlike other biometrics, acquiring face images is non-intrusive, i.e. it has no requirements for physical contact. In 1960, Woodrow W. Bledsoe negotiated with the US government for developing a first-ever semi-automated face-recognition system (FRS). The system extracts facial landmark features from photographs to do recognition. The performance of this system depends on the discriminant of the feature points being used. In 1988, researchers started searching for criminals from video sequences based on a database of mugshots. In the same year, two researchers, named Kirby and Sirovich, developed a linear algebra technique known as Principal Component Analysis (PCA) [1], and used it to solve the face recognition problem. It is considered as a major breakthrough in the face-recognition community, as face images were approximated by using less than a hundred values. In 1991, Turk and Pentland [1] found that during the usage of PCA technique, the generated residual errors could also be used for face detection in an image. This result brought new insights to the research of developing real-time FRS. After that, various programs were initiated by the US government to boost the research in facial biometrics. Some of the widely known programs include face recognition technology (FERET), face recognition grand challenge (FRGC), etc. Since then, researchers have developed highly efficient face recognition algorithms [2-8], but some challenges, which require considerable attention, still exist, such as low-resolution face recognition [9-11], age-invariant face recognition [12-14], and occluded or noisy face

recognition. This motivates us to propose some effective solutions for these challenging applications. Therefore, in this thesis, we focus on solving these three major problems by developing robust algorithms with state-of-the-art performance.

1.2 Major Challenges

Practical biometric systems are often confronted with low-resolution (LR) or poor-quality images captured by surveillance cameras, which are quite difficult to identify by the conventional face recognition systems. Due to the availability of large amounts of training data, and high computational power, face-recognition algorithms based on deep learning have achieved remarkable performances in recognizing medium resolution (MR) and high-resolution (HR) face images, captured in unconstrained environments. The highest recognition rate on one of the challenging face datasets, Labeled Faces in the Wild (LFW) [15], is more than 99 percent, which even outperforms human-level performance. However, it is still finding a way to make its mark in solving the low-resolution (LR) face recognition problem. Until now, many issues in LR face recognition have remained un-solved. They include super-resolution for face recognition, face detection from a long distance, resolution-robust features, and unified feature sub-spaces. Another major challenge is the large intra-personal variations caused by age progression. In this regard, the main challenge is to develop an efficient, discriminative feature representation and matching framework, which is robust to aging variations. Until now, only a few studies have addressed this problem. Furthermore, data available for training and testing may be corrupted by different types of noise variations, disguise or occlusion. Previously proposed methods, e.g. PCA [1] and sparse representation-based classifier (SRC) [7], did not pay attention to the possible contamination in training data, which heavily degrades their performance when corrupted testing data is presented for recognition. Therefore, there is a need to develop an accurate FRS, which is robust to possible corruption in both training and testing data. In the next chapter, we will briefly review the existing approaches for facial image analysis and recognition in both constrained and unconstrained environments. Particularly, we will also review the existing methodologies for solving the low-resolution and age-invariant face recognition problems.

1.3 Statements of Originality

This thesis claims the following contributions to be original.

a) An efficient sparse coding-based algorithm is proposed to recognize low-resolution face images under both controlled and uncontrolled environments. Different from other approaches, our method first down-samples

gallery faces to the same resolution as the query face image, and performs recognition in the low-resolution domain. Our method first extracts robust local features from face images, which are then decomposed into a low-rank feature matrix and a sparse error matrix. After that, learned low-rank part is projected onto a low-dimensional feature subspace using our proposed sparse coding-based objective function. Finally, feature matching is performed using the linear-regression model, which provides superior performance.

b) An efficient and robust deep-feature encoding-based discriminative model is proposed for solving the cross-age face recognition problem. Our algorithm first extracts deep-features from face images using a pre-trained deep Convolutional Neural Network (CNN) model (AlexNet), and then learns an age-discriminative codebook using a pair of training images with a large age gap. Finally, the gallery and query image's features are encoded using our proposed locality-constrained feature encoding framework. To speed up the encoding process, a specific number of local bases (nearest neighbours) are selected from the learned codebook. These encoded features are then fed to the linear-regression-based classifier to do face recognition.

c) To address noise variations in cross-age face images, an efficient manifold-constrained low-rank decomposition algorithm is proposed that converts extracted deep features into a low-rank feature matrix by incorporating the local structural information of the data samples. These learned low-rank features are then encoded using our proposed feature-encoding scheme based on locality information. Our algorithm first learns an age-discriminative codebook by fusing deep features from a pair of training images with a large age difference using kernel canonical correlation analysis (KCCA). In the testing stage, the learned low-rank gallery and the query image's features are encoded using a learned codebook. Finally, the NN classifier is utilized to do face recognition.

1.4 Thesis Outline

This thesis is structured into six chapters, which are organized as follows:

Chapter 2 reviews the state-of-the-art techniques related to facial image analysis and recognition under constrained and unconstrained environments. Firstly, those pioneer works in the field of face recognition are reviewed including, appearance-based and model-based face recognition methods. After that, some feature extraction and classification techniques are briefly reviewed. Some evolutionary work related to deep learning is reviewed. Previous and current research on existing challenges in the field of face recognition, such as low-resolution (LR), and aging face

recognition is briefly reviewed. Furthermore, existing state-of-the-art algorithms related to our proposed frameworks in this thesis are also reviewed.

Chapter 3 presents our sparse coding-based algorithm for recognizing the LR face images. The proposed framework first decomposes a multiple of extracted local features into a low-rank feature matrix and an associated sparse-error matrix. After that, the learned low-rank part is used to learn a projection matrix based on our proposed sparse-coding-based algorithm, which preserves the sparse structure of the learned low-rank features, in a low-dimensional feature subspace. Finally, a coefficient vector, based on linear regression, is computed to determine the similarity between the projected gallery and query image's features. Furthermore, a new morphological pre-processing approach is also proposed that aims to improve the visual quality of images. Our experiments were conducted on five available face-recognition datasets, which contain images with variations in pose, facial expressions and illumination conditions. The proposed approach provides superior performance in recognizing LR face images even of the size 8×8.

Chapter 4 presents our proposed robust deep-feature encoding-based approach for solving age-invariant face recognition problem. It is capable of recognizing face images with large age gaps, and also has proved to be robust to noise variations. The algorithm first extracts high-level deep-features from face images using a pre-trained deep CNN model (AlexNet), which are then encoded using an age-discriminative codebook. To make the codebook discriminative in terms of age progression, CCA is utilized to project pairs of training face images with large age gaps onto a coherent feature subspace, such that correlation among them is maximized. After learning a codebook, the gallery and query image's features are encoded using our proposed locality-constrained feature-encoding framework. The encoded features are then passed to the linear-regression based classifier for recognition.

Chapter 5 presents an age-invariant face recognition method based on a deep low-rank feature-learning and encoding framework. The method is also capable of handling possible corruption in the training and testing data by learning low-rank deep features, using a proposed manifold-constrained low-rank decomposition algorithm. After extracting deep features from corrupted face images, the learned features are then decomposed into a low-rank feature matrix and a sparse-error matrix by preserving the local structure of the features. The learned low-rank features are then encoded using our proposed feature encoding-based algorithm, which enhances the discriminative power of the features. Finally, the NN classifier is employed to do face recognition. Furthermore, the periocular region of a human

face is investigated in terms of age progression using the proposed framework, which also provides superior performance in terms of recognition rate.

Finally, our proposed research work is concluded in Chapter 6, along with the discussion. Furthermore, some more existing challenges and possible future research directions in the field of face recognition are also discussed.

Chapter 2 Literature Review

In this chapter, we review existing works related to facial image analysis and recognition, along with recent advancements made in this field of research. The existing methods are first classified into different categories, and their performance in both controlled and uncontrolled environments, is discussed. Moreover, we also review some of the feature-extraction, feature-encoding, manifold-learning, and similarity measurement methods, which we employed in our proposed methods in this thesis.

2.1 Face Recognition algorithms

Face recognition is a widely studied research topic, which has a lot of real-world applications, such as criminal identification, finding missing children, healthcare, etc. The main purpose of face recognition is to identify a given query face image by comparing it with the face images in a database. An automated face recognition system is first required to locate facial features, like nose, eyes, etc., and to normalize the facial geometry and appearance, like illumination. This can provide a more convenient and reliable representation in the face feature space. An appropriate selection of facial features and classification techniques are the building blocks of the FRS. The pipeline of a face recognition system involves three major steps: (1) face detection and alignment, (2) facial features extraction, and (3) feature matching. In the first stage, a face region is first detected in each given image. It is considered an important pre-processing step, so it must be able to handle face images taken under large expression, pose, and lighting variations. For face detection, existing methods can be categorized into two classes, which are feature-based methods [16-21], and model-based methods [22-25]. The second and the most crucial step of face recognition is to extract distinctive useful information from face images. Feature-extraction techniques must be robust to various geometric and noisy variations. Finally, the extracted features are fed to an appropriate feature-matching module or classifier, which compares the query image's features to the images in a gallery database, and identifies the one with the highest similarity score.

Fig. 2-1. Face images captured under different lighting conditions.

In the real-world scenario, practical face recognition systems must be able to perform well under controlled and uncontrolled environments. There are various factors that affect the face-recognition performance. These include uneven lighting conditions, pose variations, facial expressions, low-resolution, aging variations, occlusion, disguise, noise, etc. Among the first three factors, pose variation is the most challenging one, as most of the images available in the gallery set are usually of frontal view. Similarly, images taken under uneven lighting conditions are difficult to identify, as most of the facial parts cannot be seen. Fig. 2-1 shows the face images taken under different lighting conditions. With the revolution of deep learning-based methods in the face recognition research, remarkable progress has been made in addressing these three facial variations. Currently, the highest recognition rate on the challenging LFW dataset is more than 99%. However, the performance drops rapidly when the image resolution becomes lower than 25×25. Moreover, existing face recognition methods are sensitive to noise, occlusion, and aging variations. Therefore, researchers are now focusing on solving these various challenges, with the aim of developing an accurate and robust FRS. In the coming sections, we will review and analyze the performance of existing methods in addressing these kinds of challenges.

2.1.1 Appearance-based Face recognition Methods

In addition to machine learning problem, face recognition is also considered as a space searching problem. The major objective of appearance-based methods is to learn a discriminant low-dimensional subspace where data samples (face images) can be projected for classification. In this context, the earliest method was proposed in 1991, namely principal component analysis (PCA) [1], also known as Eigen Face. It uses covariance matrix of the probability distribution over high-dimensional space. It is also known as linear dimensionality reduction technique that allows few principal components to represent the training set. The generated Eigenfaces can be linearly combined to reconstruct the images in the original training set. Another classical method is linear discriminant analysis (LDA) [26], also known as Fisherfaces. LDA is a supervised subspace learning technique, which maximizes the between-class scatter matrices, while minimizing the within class scatter matrices. Furthermore, it provides more discriminant information than PCA and tends to outperform it if a large number of training samples are available. Another method proposed was independent component analysis (ICA) [27], which is a generalized version of PCA. It minimizes the higher order dependencies in the given input data and projects face images onto the basis vectors that remains independent as possible. However, these methods do not preserve the local structure information while projecting data samples onto a

new low-dimensional subspace. To solve this issue, a linear mapping technique known as locality preserving projection (LPP) [28] was proposed, which projects data in the direction of maximum variance by preserving the neighborhood structure of the images. These linear mapping methods can be good enough but does not incorporate the non-linear structural information of the data samples. To address this issue, various non-linear dimensionality reduction methods [29-32] are proposed, also known as manifold-learning. Isomap [29] is the earliest proposed manifold learning algorithm, which projects data points onto a low-dimensional subspace, such that geodesic distances among the samples are preserved. The data points are reconstructed in a new subspace using nearest neighbor and shortest path graph search. Finally, it becomes an Eigenvalue decomposition algorithm where m largest eigenvalues are selected to construct a new low-dimensional subspace. Another popular and widely used manifold learning algorithm is locally linear embedding (LLE) [30], which identify and exploit local symmetries of the data points to learn the manifold feature subspace. Instead of computing the pairwise distances, it recovers the global structure of the data from the locally linear fits. According to the assumption made by LLE, the data points and their corresponding neighbours have a linear relation in the manifold subspace, so neighboring points can be used to reconstruct each data point. Some other state-of-the-art manifold learning algorithms include Multidimensional scaling [31], Laplacian Eigen mapping [32], etc. These methods enhance the discriminative power of the input data points, and provides much better performance than linear-mapping-based methods.

2.1.2 Model-based Face Recognition Methods

To learn the variations in facial expressions, model-based methods were proposed that constructs 2D and 3D models of a human face image. It utilizes the prior information of the human face to design the model. Wiskott et al. [33] developed a feature-based model, based on elastic bunch graph. The method represents face images using a labeled graph, based on Gabor wavelets. A simple similarity function is then used to compare the graphs of new faces. Furthermore, constructed graphs are capable of handling rotational variance in depth. Cootes et al. [24] proposed a 2D morphable face model by utilizing both shape and texture information. The model can be generalized to any valid sample. In the training stage, the relationship is learned between the residual errors and the parameter displacements, which exist between a training sample and a synthesized image. The method is capable of performing efficient matching with only a small number of iterations.

There are various advantages and disadvantages of model-based methods because of their physical relationship with real faces. Extracting facial feature points with robustness is a challenging task that strongly depends on the model fitting. Inaccurate fitting can reduce recognition accuracy up to a considerable level.

2.2 Feature Extraction Techniques

Performance of face recognition systems heavily depends on type of discriminant information extracted from facial images. In the last two decades, several feature extraction methods have been developed, which provides superior performance in many machine learning tasks, e.g. object detection, image classification, face recognition, etc. Features can be categorized into two classes, known as global features and local features. For recognition, local features have been proven to be more robust to variations in facial expressions, pose and illumination as compared to global features. Most commonly used local feature extraction methods include 2-D Gabor wavelets (GWs) [34, 35], Local binary pattern (LBP) [2], Histogram of oriented gradients (HOG) [36], scale invariant feature transform (SIFT) [37], Discrete cosine transform (DCT) [38], etc. Among these, GWs and LBP are the widely used features for face recognition. Gabor wavelets is considered as a good choice for performing space-frequency localization, and provide optimum resolution in both spatial and frequency domains. This makes it robust against various facial variations. LBP was originally proposed for texture classification having very less computational complexity. It works by comparing the given central pixel with the neighboring ones, and computes the binary code. If the value of central pixel is greater than the neighboring pixel then it assigns the value '1', otherwise '0'. Later, it was used for face recognition, and outperforms Gabor wavelets. Moreover, LBP is insensitive to monotonic gray level transformations, and performs well under different lighting conditions. SIFT is a widely used local feature descriptor for object recognition. First, it detects the key points in an image using difference of Gaussian (DOG) function, and then computes HOG at each detected key point. This feature has been proved to be invariant to various geometric transformations, such as translation, rotation, and scaling of the data samples. The faster version of SIFT is known as Dense SIFT (DSIFT) [39]. DSIFT excludes the step of key point detection, and extracts local features at every pixel of an image, which allows DSIFT to get more discriminative information for recognition as compared to SIFT. These feature descriptors can be combined with other recognition frameworks to achieve better performance.

Recognizing face images under uncontrolled conditions is a challenging task, and no single feature is good enough to tackle all the facial variations simultaneously. Therefore, combining multiple features is a promising way to improve

the recognition accuracy. There are three different ways to combine the extracted information, which are: (1) Feature-level fusion, (2) Decision-level fusion, and (3) matching score-level fusion. One major example of feature-level fusion was proposed in [40], where Gabor and LBP features are fused to get more discriminative information. LBP has the capability to extract even small appearance details, while GWs can extract shape information at multiple scales and orientation. Their complementary nature makes them promising candidates for feature fusion. Furthermore, fusion of these two feature sets produces a high-dimensional feature vector, so PCA [1] can be used to reduce the feature dimension. For decision-level fusion, different classification techniques are first selected that makes their own decisions. Finally, decisions from multiple classifiers are integrated to produce the final decision. For match score-level fusion, scores computed using different biometric matchers are fused to get the final decision. It is worth noting that the matching scores generated by different matchers may have a large difference. The scores generated by different modalities can be combined to get a final similarity score.

2.3 Classification techniques

The final and the most crucial step of face recognition system is classification. There exist several methods for classification that includes k-Nearest Neighbour (kNN) [41], Support Vector Machine (SVM) [42], artificial neural network (ANN), sparse representation-based classifier (SRC) [7], Linear and logistic regression [43], Decision trees [44], etc. It is observed that linear classifiers cannot provide satisfactory performance, if features are not linearly separable. Classification techniques can be categorized as either supervised, unsupervised or semi-supervised. Most of the face recognition systems use supervised learning methods, e.g. *k*NN classifier that uses class label information and assigns a given sample to the class to which the majority of its k-nearest neighbors belongs to. SVM is a widely used classifier, which ensures the maximum distance between the hyperplane and the points near to the decision boundary. It is a supervised learning technique that produces an optimal hyperplane to classify new samples. There are three main ideas in building a good classifier which are similarity, decision boundary, and probability. Similarity concept is quite simple in which metric is established to define and represent the similarity between the images of the same class. In this regard, several metrics have been proposed including Euclidean distance, Chi-square, Hamming distance, etc. Some classifiers are based on probabilistic approach which assigned patterns to the class with the highest estimated posterior probability. Bayesian classifier [45-46], multi-layer perceptron (MLP) [47] (trained under a suitable loss function), and logistic regression are one of the major examples of probabilistic classifiers. Naïve Bayesian classifier

applies Bayes theorem with strong assumptions between the given features. It is widely used for document or text classification by using word frequencies as features, where the parameters are determined by using maximum-likelihood. MLP is a class of ANN that utilizes a backpropagation technique for training. It is different from a linear perceptron due to multiple layers and non-linear activation functions. Furthermore, it has a capability to classify the data samples which are not linearly separable.

In decision-boundary based classifiers, the measurement error between the training and testing samples is minimized. Fisher discriminant analysis (FDA) [26] is one of the examples that is used to model the differences between the classes of data, and then minimize the mean square error. SRC [7] is a state-of-the-art classifier proposed for robust face recognition. According to sparse theory, each test sample can be linearly represented in terms of all training samples. It works by first computing the sparse coefficients using l_1 minimization technique, and then computes the residual value. Finally, given input sample is classified based on the least residual value. Later, we will briefly discuss the concept of sparse coding for face recognition. Naseem et al. [43] proposed a linear regression concept for classifying face images. It assumes that there exists a linear relationship between a probe image and all the samples in a gallery set. If a query face image fits to the i^{th} class in the gallery set, it can be represented as a linear combination of the gallery-images features from the same class. The relationship between a probe image and a gallery image is determined using least-squares method, and a probe image is assigned to the class with the minimum reconstruction error.

2.4 Deep learning-based Methods

Deep-learning models have revolutionized pattern-recognition research by providing extra-ordinary performances, which is quite close to human-level performances. One of the main reasons for its success is the availability of a large training sets and the networks are trained for feature extraction and recognition from end to end. For face recognition, various deep learning (DL) models [3, 6, 48-54] have been established, and provide excellent performance. Sun et al. [3] proposed a deep-CNN model, namely DeepID, which learns high-level deep features from the patches of face regions for identification. Another deep face model [6], namely FaceNet, which uses a large network trained by distance constraints, was proposed. This model achieves a very high recognition accuracy of 99.60% on the challenging LFW dataset. Parkhi et al. [48] proposed a very deep-CNN architecture, namely VGG-Face, trained on 2.6M images from 2,622 identities. The model has proven to be highly successful for face recognition, and achieves 99.13% recognition

accuracy on the LFW dataset. Wen et al. [50] proposed a deep network, which minimize the intra-class distances between the deep-features in a loss function, and achieves 99.2% recognition accuracy on the LFW dataset. Recently, Liu et al. [49] proposed an angular softmax loss function, which learns discriminative features using CNN based on the ResNet architecture [55]. The method achieves 99.42% accuracy on the LFW dataset. It is to be noted that the scale of training data used by SphereFace [49] is smaller than the other deep-learning methods. Although, deep-learning based methods provides powerful data representations, the method in [56] argues that deep-learning performance is affected by various facial variations especially pose variation. Therefore, preprocessing of data samples is necessary. For deep-learning, data preprocessing can be categorized into two categories: (1) Many-to-one normalization, and (2) One-to-many augmentation. In the first category, non-frontal face images are converted into frontal ones by using face-frontalization method, and then FR is performed. However, in the second category, one image is used to generate multiple images with different poses. These images are then used to train deep neural networks, which enable it to learn pose-invariant features. The commonly used data-augmentation methods include various geometric and photometric transformations, such as mirroring, oversampling, and rotation of the face images. By doing this, DL models can learn a rich level of feature representations, which further improves the recognition rate. To meet the requirements in terms of large training data, several datasets have been released to train Deep-CNN models for recognizing face images, e.g. CASIA-Web Face [57], Mega Face [58], etc. CASIA-Web face contains 494,414 face images from 10,575 identities. In 2015, the largest dataset, namely, Mega Face was released to check the performance of existing face recognition methods. It consists of 1 million distractors from more than 690K identities. The training set consists of 4.7 million photos from 672,057 identities, while the testing set contains both images of celebrities and non-celebrities from the FaceScrub and FGNET datasets, respectively.

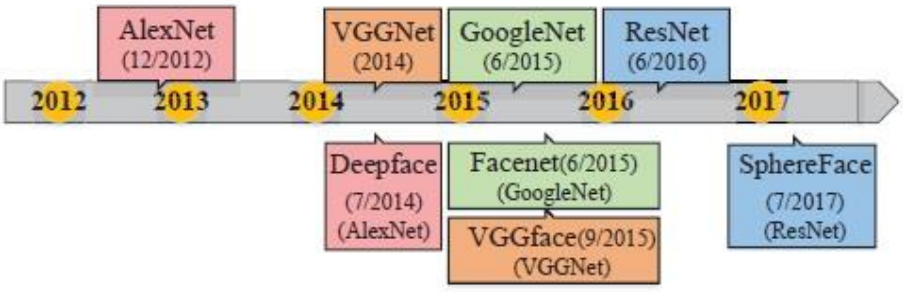


Fig. 2-2. Some state-of-the-art deep architectures proposed for image classification, and successfully utilized for face recognition. (Image adapted from [59]).

Loss Function: Loss functions calculate the loss between the output and the target variable, which plays a major role in training a deep neural network. The commonly used loss function in deep-CNN architectures is softmax loss function. However, due to large intra-personal variations in face recognition, softmax loss function becomes ineffective. Researchers are now focusing on developing more novel loss functions, which makes the learned features more discriminative. Some of the recently proposed loss functions include: (1) Euclidean-distance-based loss function, which increases the discriminative power of the features by minimizing the intra-class-variance and maximizing inter-class-variance using Euclidean distance. (2) Cosine/Angular loss function [49], which enhances the discriminability of deep features by learning an angular similarity. Moreover, the performance of loss function can be enhanced by using L2 normalization. Once the deep-CNN models are trained using large amount of data and reasonable loss function, a given query input can be passed to the trained deep network to extract high-level features. After extracting the features, many metric learning techniques can be used to measure the similarity score.

2.5 Review on Existing Challenges

2.5.1 Low-resolution Face Recognition

Face images captured by surveillance cameras are usually of low-resolution (LR) and poor-quality, with huge variations in pose, facial expressions and lighting conditions, as shown in Fig. 2-3. These make the low-resolution (LR) face recognition task very challenging. Conventional face recognition approaches [1, 26, 27] gives good performance when captured images are of high-resolution and taken under controlled environments, but their performance degrades heavily when the image resolution becomes lower than 25×25. A lot of research is being conducted to tackle these challenges separately, and handling all of them simultaneously is very challenging and requires significant attention.

Fig. 2-3 Low-resolution face recognition (Image adapted from paper [193])

There are three possible approaches for solving LR face recognition problem as shown in Fig. 2-4. One of the possible solutions is to enhance the visual quality of an input LR image using a super-resolution (SR) technique, and then perform recognition. Second approach is based on coupled mapping in which HR gallery image and LR query image are projected onto a common subspace, where recognition is performed. Third approach is to first down-sample the HR gallery image to the same resolution as LR query image, and perform recognition in the LR domain. In this section, we will discuss some possible solutions for LR face recognition along with their advantages and disadvantages.

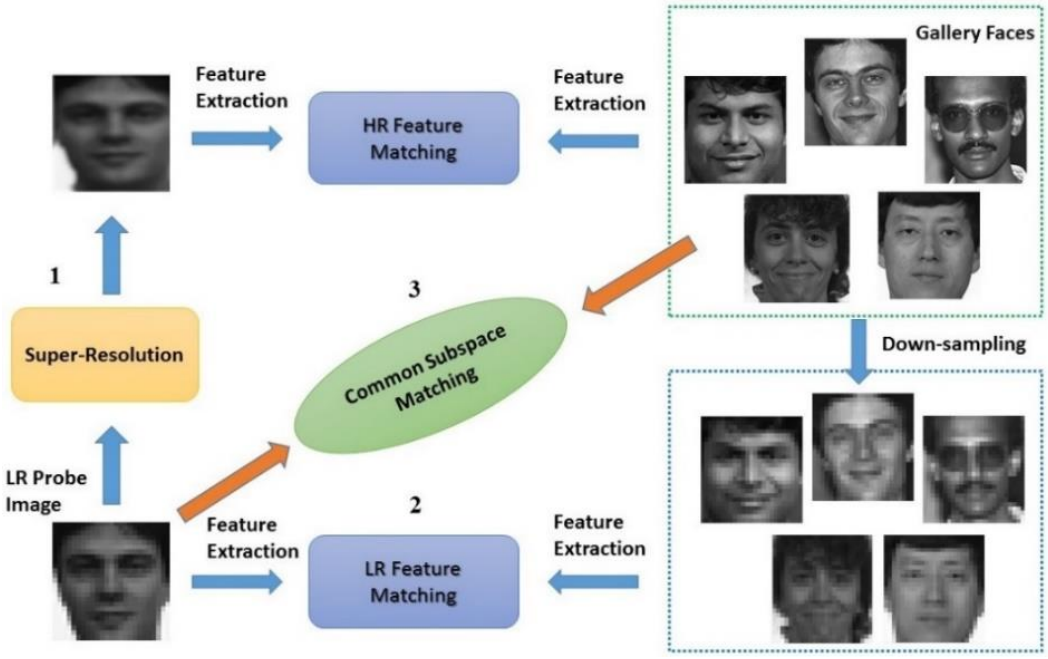


Fig. 2-4: Three possible approaches for solving LR face recognition.

2.5.1.1 LR Face recognition using Super-resolution techniques

For face images, the SR process is also known as face hallucination, which was first proposed in [60]. The method decomposes a face image into a pyramid of features by utilizing the Gaussian and Laplacian pyramids, and then reconstructs the corresponding high-resolution (HR) image. In [61], the limitations of SR were discussed and some possible solutions to breaking them were given. Yang et al. [62] proposed a SR approach based on sparse coding, which generates the HR image patch by computing the sparse representation coefficients of each LR image patch from a dictionary. Finally, the dictionaries of both HR and LR image patches are trained simultaneously to enhance the similarity between the HR and LR image pairs. Wang et al. [63] utilizes principal component analysis (PCA) to linearly represent LR test image, in terms of similar LR training images. The HR image is constructed, by replacing the LR

training images with the corresponding HR training images. The SR methods in [64, 65] assume that the LR and HR images of the same person have some intrinsic correlation. Another face-hallucination framework presented in [66] assumes that two face images of the same identity have high-correlation in terms of their local-pixel structures. The approach learns the local-pixel structure for reconstructing a HR image by searching a face database for similar HR faces using the LR input image. It was reported in [10, 67] that super-resolved images contain distortion and artifacts, which reduce the recognition accuracy, and hence are not a feasible solution to LR face recognition. Hennings-Yeomans et al. [68] proposed an objective function for performing hallucination and recognition simultaneously. This approach proves to be computationally expensive, because optimization is required for each test image. Huang et al. [69] proposed a SR method, which performs non-linear mappings of coherent features. The method learns a coherent subspace between the HR samples and LR samples using canonical correlation analysis (CCA). Radial basis function (RBF) is then used to learn the nonlinear mappings between the coherent features, and the super-resolved coherent features of a LR image are determined by using a trained RBF model. Zou et al. [9] proposed a framework for face hallucination, which aims to learn a mapping function that defines the relationship between the HR and LR image spaces by utilizing a new discriminative constraint. Jian et al. [70] proposed an improved method for performing hallucination and recognition simultaneously by utilizing the singular value properties of images at multiple resolutions. To recognize LR face images, only the super-resolved HR features are required. One way is to extract features from super-resolved HR images. However, these images, which are distorted versions of the true HR face images, are generated by estimation. Pong et al. in [71] proposed an approach, which directly estimates the HR features for recognition by performing super-resolution in the feature space. The method also fuses the features from different resolutions, so as to further improve the recognition accuracy.

2.5.1.2 Coupled-Mapping Based Methods

Another approach for recognizing LR face images is based on coupled mappings. Li et al. [10] proposed to learn a unified low-dimensional feature subspace for LR and HR images, which then facilitates the ultimate classification. Zhou et al. [72] presented a method, which preserves the discriminative power of the HR and LR samples in the learned common subspace using simultaneous discriminant analysis. Ren et al. [67] learned a common subspace for LR and HR samples using coupled kernel embedding, which uses a new similarity measure to compare the multimodal data. Biswas et al. [73] used multidimensional scaling for LR face recognition, which projects LR and HR samples into a

common subspace, such that the geometrical structure of the samples is preserved. This approach also ensures that the distance between the two LR images is nearly the same as that of its HR counterparts. Siena et al. [74] utilizes the local structure's relationship between the HR gallery images and LR probe images to learn a common subspace. Similarly, Shi et al. [75] projects HR gallery images and LR probe images into a unified latent subspace by incorporating the geometrical structure of each given sample with respect to its neighboring points. The approach combines all the local optimizations to construct a global structure, which preserves the discriminant information of the samples in the learned subspace. Zhang et al. [76] proposed to learn a projection matrix, which maximizes the margin between inter-class and intra-class distances in the common subspace. Wang et al. [77] used CCA to determine the correlation between HR and LR image pairs, such that a pair of transform matrices are computed for the HR and LR face images, respectively. Jiang et al. [78] addressed the LR face recognition problem by proposing a coupled discriminant method based on multi-manifold analysis. The approach learns the local structure as well as neighborhood information about the manifold subspace covered by the image samples. After that, two mapping functions are learned to project the HR and LR samples, respectively, into a common feature subspace. Chu et al. [79] proposed a cluster-based method, based on simultaneous discriminant analysis. The method learns the cluster-based scatter matrices to regularize the between-class and with-in class scatter matrices. This enhances the discriminability of the feature space. Xing et al. [80] proposed a coupled mappings-based approach, which projects face-image samples into a unified feature subspace using the topology of a generalized bipartite graph. The approach also preserves the local geometrical structure of the samples when they are projected into a new subspace. Recently, Yang et al. [81] proposed a discriminative multi-dimensional scaling (DMDS) method for LR face recognition, which considers the intra-class, as well as inter-class, distances, while projecting the HR and LR data samples into a unified feature subspace.

2.5.2 Age-invariant Face Recognition

Recognition of face images under large age-variations is a challenging research problem, which receives considerable amount of attention in the past few years. It has many practical applications, e.g. criminal identification using photographs, finding missing children, etc. The major challenge in AIFR is the considerable intra-personal variations due to age progression as shown in Fig. 2-5. Due to the availability of large amount of training data and computational power, existing deep-learning-based techniques have already achieved superior performances in recognizing face images under unconstrained environments. However, their performance is limited in solving aging

face recognition problem. There are some reasons for that. Firstly, it is quite difficult to collect training data with large age differences as it requires long period of time and great effort. Secondly, modeling age progression is challenging as it is specific to different persons. For example, some people use different kinds of make-up and skin-refreshment techniques to look like younger people. On the other hand, some people go through a lot of stress and problems in their life, which makes them look like elder people at the younger age. Moreover, photos taken at earlier ages are of poor quality and contains a lot of distortions, as the high-resolution cameras were not available at that time.

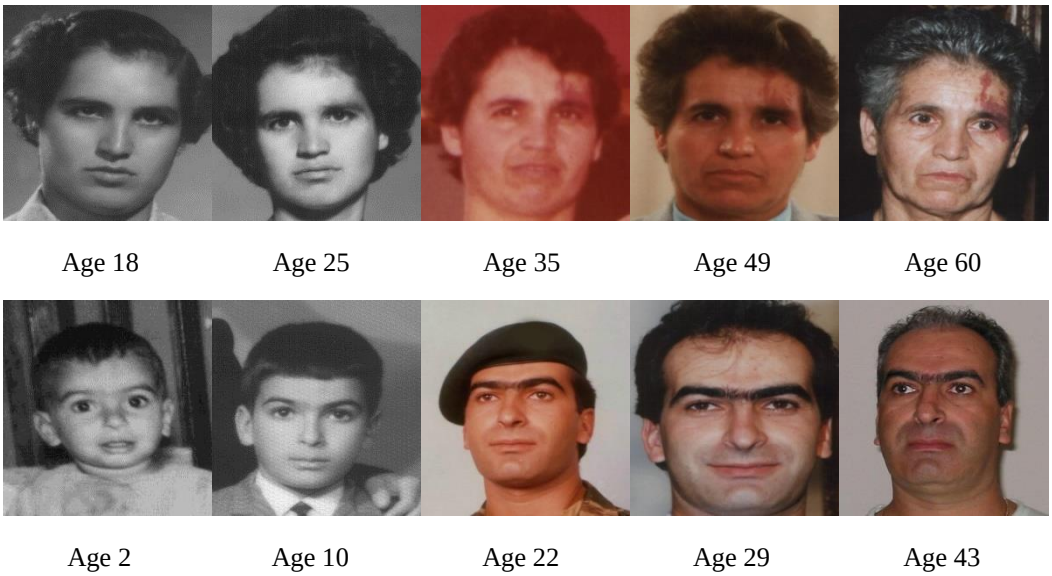


Fig. 2-5. Sample face images from the FGNET dataset from two different people with large age variation, where each row represents the face images of the same person.

2.5.2.1 Generative Models

Existing research work related to facial aging mainly focuses on either age estimation [82-88] or age simulation [89-96]. A typical approach to AIFR is to synthesize a test face image to be the same age as the gallery image before performing recognition, i.e. it is based on generative models. Lanitis et al. [87] developed a statistical model to encode face images in a compact manner. The approach utilizes training images to establish the relationship between encoded features and the real ages of the subjects in the corresponding images. Based on this relationship, the age of an unseen subject can be estimated. Ramanathan et al. [88] proposed an age-invariant face verification framework by proposing a fast-growing model for people whose ages are under 18 years old. Park et al. [91] proposed to compensate for the aging variation by developing a 3D aging model using a 2D face-aging dataset, which enhances the recognition performance. Suo et al. [92] developed a dynamic method for modeling human age progression. The approach represents face images in different age groups by using a hierarchical graph. Wang et al. [96] developed an aging

simulation method by transforming the shape and texture of a human face from its source age to the target age, which is performed in the eigenspace.

These generative model-based methods have some limitations. First, it is quite difficult to build a face model as it cannot represent the aging process in an accurate way, especially when the training set is small. Constructing an accurate aging model requires strong parametric assumptions, as well as the real ages of the training images, which makes it unsuitable for real-time face recognition. Furthermore, the training data must be taken under controlled environments e.g., frontal pose, neutral expression, and normal lighting.

2.5.2.2 Discriminative Models

Recently, discriminative models have been proposed for AIFR, which focus on developing features robust to aging variations. Ling et al. [97] utilized the gradient orientation pyramid (GOP) for feature extraction, and the support vector machine (SVM) for classification. Guo et al. [98] investigated the effects of age gaps on recognition rates by performing experiments on a large data set, using PCA. In [14], the feature descriptors, multi-scale local binary pattern (MLBP) and scale invariant feature transform (SIFT), were utilized, and a fusion framework, based on random sampling, was proposed to enhance the recognition performance. Extended versions of random sampling, methods based on local discriminant analysis (LDA) were also utilized in [99, 100] to address the aging variations in face recognition task. These approaches have been proved to be robust with only few requirements regarding associated parameters and training data. Gong et al. [101] proposed a novel feature descriptor, namely maximum entropy feature descriptor (MEFA), which analyses facial images on a micro level and encodes the information in a form of discrete codes. The approach is based on a new matching framework, known as identity factor analysis, which estimates the probability that the two given face images are from the same identity.

The recent work on aging face recognition in [102] proposed a probabilistic model, which represents a face image with two components, namely the identity component and the aging component, respectively. A learning algorithm, which estimates these two components simultaneously, was proposed. Chen et al. [103] proposed a new coding framework, which performs feature encoding by using reference images with an age-invariant reference space. It is based on the assumption that if two persons look similar at younger ages then they might look alike at the older ages. By utilizing the large size face-aging dataset as a reference set, the method learns age-invariant reference space for feature encoding. The approach proves to be highly scalable as it only needs a linear projection for feature encoding in

the testing stage. In [104], a joint additive model was proposed to perform cross-age face verification, which is solved by the alternating greedy coordinate descent (AGCD) algorithm. Recently, a hierarchical model with two stages of learning is proposed in [105]. Firstly, discriminative features are learned from microstructures of facial images, which are then converted into integer codes for face recognition. After that, the extracted information is polished by using the output obtained from the first stage. However, these methods utilize hand-crafted features, and their performances strongly depend on the properly pre-processed face images.

2.5.2.3 Deep-learning based Approaches for AIFR

Wen et al. [106] learned age-invariant features by proposing a deep convolutional neural network based on latent identity analysis. The method keeps the identity components separated from the aging variations in the learned deep features. The model's parameters are then used to update the parameters of a latent-factor fully connected layer. The latent identity model and the corresponding loss function invariant to aging variation guides the learning process of the proposed convolutional neural network (CNN). However, it assumes that the combination of identity and aging features is linear and can be separated completely. Therefore, the method ignores the probability regarding the correlation between the aging and identity component. Another deep-learning framework [107], for age estimation, was proposed to learn age-invariant features. The method utilizes the real-age and identity labels of the training data to simultaneously perform age-estimation and face recognition. Recently, another deep learning framework [108] is proposed, which combines two networks consists of identity and aging information, respectively. Furthermore, both networks share the same feature layers. Cross-age identity features are separated from aging features by training the fused networks alternatively. Xu et al. [109] proposed to learn the complex non-linear aging progression using an auto-encoder network. The method decomposes a face image into identity, age, and noise component by proposing a non-linear factor analysis method. Recently, a novel distance metric optimization method [110] based on deep learning was proposed. The method used a large number of matched pairs from the training set to learn identification information. These matched pairs are then served as an input to enhance the differences between the unmatched pairs. The model parameters are updated using the classical gradient descent algorithm. The learned features and distance metric are optimized simultaneously. Wang et al. [111] proposed a deep CNN model, which learns age-invariant deep features. The learned deep features are decomposed into orthogonal identity and aging components. The identity components are then used for face recognition.

2.6 Review on Related Methods

In this section, we will review some of the existing methods that are closely related to our proposed methodologies in this thesis. These methods include linear and nonlinear subspace learning methods, sparse representation theory, feature-encoding methods, low-rank feature learning, and analysis of the periocular region for face recognition.

2.6.1 Principal component Analysis

Principal component analysis (PCA) is an unsupervised linear dimensionality reduction technique, which uses only few basis vectors to represent the set of images. It is a linear mapping technique that projects high-dimensional data onto a low-dimensional subspace by retaining maximum amount of information. Let us consider that the high-dimensional space is represented by $x = a_1 v_1 + a_2 v_2 + \cdots + a_N v_N$, where $v_1, v_2, \dots, v_N$ is the basis of the N-dimensional space. Similarly, low-dimensional space is represented by $\hat{x} = b_1 u_1 + b_2 u_2 + \cdots + b_K u_K$, where $u_1, u_2, \dots, u_K$ is the basis of the K-dimensional space. PCA works by minimizing the error $\|x - \hat{x}\|$. The optimum low-dimensional subspace is determined by computing the few eigenvectors of the covariance matrix of x. These eigenvectors are associated with the corresponding eigenvalues, which are also known as principal components. Suppose we are given with N-dimensional vectors $x_1, x_2, \dots, x_M$, where M represents the total number of vectors. First step is to compute the mean, which is defined as $\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$. The second step is to subtract each sample from the mean defined as $\varphi_i = x_i - \bar{x}$. The covariance matrix of the data sample can be computed as:

$$\boldsymbol{C} = \frac{1}{M}\sum_{n=1}^{M} \varphi_n \, {\varphi_n}^T = \frac{1}{M}\boldsymbol{A}\boldsymbol{A}^T, \tag{2.1}$$

where $\boldsymbol{A} = [\varphi_1, \varphi_2, \dots, \varphi_M] \in \mathbb{R}^{M \times N}$. Finally, it turns out to be an eigenvector problem of the form $\lambda u = \boldsymbol{C}u$, where u are the eigenvectors of covariance matrix $\boldsymbol{C}$, and λ are the associated eigenvalues. The first K eigenvectors having largest eigenvalues are selected as the corresponding basis vectors to construct the low-dimensional space. In face recognition community, it is also known as Eigenfaces. From geometric point of view, PCA performs linear mapping of data samples in the direction of maximum variance. These directions are represented by the corresponding eigenvectors and eigenvalues. Furthermore, it preserves the global structure of the data samples, while projecting the data samples onto the low-dimensional subspace, and performs well in recognizing face images under large variations in facial expressions. However, it is unable to perform well under large pose and illumination variations.

2.6.2 Locally Linear Embedding (LLE)

The local features extracted from face images contain rich amount of information, which is usually of high-dimension. Since the 1990s, many subspace learning methods have been developed, that aims to learn a discriminative low-dimensional feature subspace, which reduces the computational complexity of the whole system. Most of these methods originated from the literature [1], where PCA was applied to represent the human face images. Another classic subspace method is LDA [26], which uses a smaller set of basis images to project data samples onto a new low-dimensional subspace, according to their class labels. However, PCA and LDA cannot incorporate the local structure of the data samples accurately, which makes them unsuitable for face recognition in unconstrained environments. Previously proposed methods, including multidimensional scaling (MDS) [31], which learns a low-dimensional subspace, such that the pairwise distances among the data points are preserved. Instead of computing the pairwise distances, LLE recovers the global structure of the data from the locally linear fits. It assumes that the data points and their corresponding neighbours have a linear relation in the manifold, so the neighbouring points can be used to construct each data point in a new low-dimensional space. The cost function used to compute the reconstruction error can be written as follows:

$$\in (\boldsymbol{W}) = \sum_{i} \left\| \boldsymbol{F}_i - \sum_{k} \boldsymbol{W}_{ik} \boldsymbol{F}_k \right\|^2 , \tag{2.2}$$

$$\text{s.t.} \sum_{k} \boldsymbol{W}_{ik} = 1 ,$$

where $\boldsymbol{F}_i$ is the real-valued feature vector. $\boldsymbol{W}_{ik}$ indicates the weight contribution of the k^{th} data point to reconstruct the i^{th} sample. It also follows the constraint that the sum of the weights in each row of the weight matrix is equal to one. According to the principle discussed above, the weights $\boldsymbol{W}_{ik}$, which performs reconstruction of the i^{th} data point in the Q-dimensional space, can also reconstruct the corresponding data points in the q-dimensional space, where $Q > q$. In this way, LLE projects the data, such that the local structural information of each data point is preserved in the new subspace. Finally, each feature $\boldsymbol{F}_i$ is mapped into a new low-dimensional vector $\boldsymbol{V}_i$, which represents the internal coordinates on the manifold. The objective function can be defined as follows:

$$\varphi(\boldsymbol{V}) = \sum_{i} \left\| \boldsymbol{V}_i - \sum_{k} \boldsymbol{W}_{ik} \boldsymbol{V}_k \right\|^2 . \tag{2.3}$$

According to the formulation in [30], the projection matrix learned from the computed weight matrix is given as $\boldsymbol{Z} = (1 - \boldsymbol{W})(1 - \boldsymbol{W})^T$. LLE is simple and computationally efficient, compared to other manifold-learning techniques, as its optimization process does not involve a local minimum. Some more important geometrical properties of LLE are as follows:

- LLE attempts to construct the graphical representation of data points. In this context, it is similar to ISOMAP [29], but it is more insensitive to short-circuiting (overlapping of the samples belong to the different classes), as compared to ISOMAP. If short-circuiting happens, then only a few local properties will be affected. Furthermore, LLE results in the successful embedding of manifolds (non-convex in nature), which is due to the preservation of local structures.
- LLE assumes that a manifold is locally linear, so it fits a hyper-plane by using the data points and it's corresponding nearest neighbours. This assumption leads to the reconstruction weights, and makes it invariant to rotation, rescaling, and translation.
- LLE is computationally inexpensive, as its optimization process does not involve a local minimum.

2.6.3 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a useful technique to study the relationship between the two variables. It works by computing the basis vectors for both variables and projects them onto a new coherent feature subspace, such that their correlation is maximized. Given two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ with columns representing the sets of variables $\boldsymbol{x}$ and $\boldsymbol{y}$ having zero mean and unit variance. Let us consider that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the pairs of direction matrices for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The respective projection coefficients for $\boldsymbol{X}$ and $\boldsymbol{Y}$ are denoted as $\boldsymbol{U}$ and $\boldsymbol{V}$, such that $\boldsymbol{U} = \boldsymbol{\alpha}^T.\boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{\beta}^T.\boldsymbol{Y}$. It works by maximizing the following function:

$$K(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{C}_{XY} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{C}_{XX} \boldsymbol{\alpha}.\, \boldsymbol{\beta}^T \boldsymbol{C}_{YY} \boldsymbol{\beta}}}, \tag{2.4}$$

where $\boldsymbol{C}_{XX}$ and $\boldsymbol{C}_{YY}$ are the within-set covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, and $\boldsymbol{C}_{XY}$ is the cross-variance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. It can be shown that the learned direction matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the eigenvectors of ${\boldsymbol{C}_{XX}}^{-1}\boldsymbol{C}_{XY}{\boldsymbol{C}_{YY}}^{-1}{\boldsymbol{C}_{XY}}^T$ and ${\boldsymbol{C}_{YY}}^{-1}{\boldsymbol{C}_{XY}}^T{\boldsymbol{C}_{XX}}^{-1}\boldsymbol{C}_{XY}$, respectively.

2.6.4 Kernel Canonical Correlation Analysis

To study the non-linear relationship among the data samples, an extension of CCA known as Kernel canonical correlation analysis (KCCA) was proposed. As discussed above, CCA seeks a linear transformation for set of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, such that the projected features in the transformed space have maximum correlation. The major drawback of CCA is that it cannot capture the nonlinear relations between the two subjects, and linear relationship cannot always be adequate for determining the correlation between the two subjects. Firstly, the data samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are mapped onto the high-dimensional subspace by their corresponding mapping functions φ_x, and φ_y, respectively. It can be written as:

$$\varphi: \boldsymbol{x} = (x_1, \dots, x_m) \rightarrow \varphi(\boldsymbol{x}) = \left(\varphi_1(\boldsymbol{x}), \dots, \varphi_N(\boldsymbol{x})\right) \ (m < N) \tag{2.5}$$

$$\varphi: \boldsymbol{y} = (y_1, \dots, y_m) \rightarrow \varphi(\boldsymbol{y}) = \left(\varphi_1(\boldsymbol{y}), \dots, \varphi_N(\boldsymbol{y})\right) \ (m < N) \tag{2.6}$$

After mapping $\boldsymbol{X}$ using φ_x, and $\boldsymbol{Y}$ using φ_y, linear CCA is applied, which moves it from primary to dual representation. A kernel can be defined as a function K, such that for all $\boldsymbol{x}, \boldsymbol{z} \in \boldsymbol{X}$, $K(\boldsymbol{x}, \boldsymbol{z}) = \langle \varphi(\boldsymbol{x}). \varphi(\boldsymbol{z}) \rangle$. In linear CCA, the overall covariance matrix $\boldsymbol{C}$ of $\boldsymbol{x}$ and $\boldsymbol{y}$ is given as:

$$\boldsymbol{C} = E\left[\binom{x}{y}(\boldsymbol{x}^T \ \boldsymbol{y}^T)\right] = \begin{bmatrix} \boldsymbol{C}_{xx} & \boldsymbol{C}_{xy} \\ \boldsymbol{C}_{xy}{}^T & \boldsymbol{C}_{yy} \end{bmatrix} \tag{2.7}$$

We can rewrite the covariance matrix $\boldsymbol{C}$ as $\boldsymbol{C}_{xx} = \boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{C}_{xy} = \boldsymbol{X}'\boldsymbol{Y}$. The projection of data onto the direction $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be defined as $\boldsymbol{w}_x = \boldsymbol{X}'\boldsymbol{\alpha}$ and $\boldsymbol{w}_y = \boldsymbol{Y}'\boldsymbol{\beta}$, respectively. Recalling Linear CCA, function needs to be maximized can be defined as:

$$\rho = \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x{}' \boldsymbol{C}_{xy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x{}' \boldsymbol{C}_{xx} \boldsymbol{w}_x \boldsymbol{w}_y{}' \boldsymbol{C}_{yy} \boldsymbol{w}_y}} \tag{2.8}$$

Put $\boldsymbol{w}_x = \boldsymbol{X}'\boldsymbol{\alpha}$ and $\boldsymbol{w}_y = \boldsymbol{Y}'\boldsymbol{\beta}$ in (2.8), we get

$$\rho = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{\alpha}. \boldsymbol{\beta}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{Y}\boldsymbol{Y}'\boldsymbol{\beta}}} . \tag{2.9}$$

Let $K_x = \boldsymbol{X}\boldsymbol{X}'$ and $K_y = \boldsymbol{Y}\boldsymbol{Y}'$ are the corresponding kernel matrices. Substitute into equation (2.9)

$$\rho = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{\boldsymbol{\alpha}' K_x K_y \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}' {K_x}^2 \boldsymbol{\alpha}.\, \boldsymbol{\beta}' {K_y}^2 \boldsymbol{\beta}}} \ . \tag{2.10}$$

Equation (2.10) is not affected by the rescaling of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Therefore, it can be maximized subject to $\boldsymbol{\alpha}' {K_x}^2 \boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}' {K_y}^2 \boldsymbol{\beta} = 1$, respectively. This can be solved using Lagrange method, which gives the corresponding equation as:

$$L(\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}' K_x K_y \boldsymbol{\beta} - \frac{\lambda_\alpha}{2}\left(\boldsymbol{\alpha}' {K_x}^2 \boldsymbol{\alpha} - 1\right) - \frac{\lambda_\beta}{2}\left(\boldsymbol{\beta}' {K_y}^2 \boldsymbol{\beta} - 1\right) . \tag{2.11}$$

To solve equation (2.11), the derivative is taken with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and set to zero, which gives the following equation:

$$K_x K_y {K_y}^{-1} K_x \boldsymbol{\alpha} - \lambda^2 K_x K_x \boldsymbol{\alpha} = 0 \ . \tag{2.12}$$

Finally, it becomes an Eigen-value problem of the form $\boldsymbol{Ax} = \lambda \boldsymbol{x}$.

2.7 Feature-encoding-based Methods

The pipeline of any object recognition tasks contains three major steps: (1) feature extraction, (2) feature encoding, and (3) classification. Feature encoding aims to represent the extracted features in a form of visual codewords, which proves to be quite helpful in improving the discriminative power of the extracted features. The baseline approaches [112-114] were proposed to compute the histogram of visual codewords. Further advancements were made in this field by introducing different kinds of encoding constraints that preserves the structural information of the extracted local features. These types of methods are categorized into two types: (1) Feature representation using visual codewords, and (2) Computing the difference between the extracted features and visual codewords. Two issues which were normally addressed in this area are the memory consumption and computational time. Any feature encoding framework consists of two major steps: (1) codebook learning, and (2) Feature representation in terms of visual codewords using learned codebook. To understand the steps of codebook generation, we will first review two clustering-based techniques, which are commonly used for codebook initialization.

2.7.1 K-Means Clustering

Most of the feature encoding methods first partitions the local feature descriptors into different clusters (regions), such that the internal structure can be parametrized linearly. Such informative regions are known as visual codewords,

and their combination is called visual vocabulary. The most widely used technique to construct this kind of vocabulary is k-means clustering [115]. Given M training descriptors $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_M \in \mathbb{R}^D$. The k-means algorithm partitions the training descriptors into K non-overlapping segments $\boldsymbol{S}_i$, such that sum-of-squares criterion is minimized. It is defined as follows:

$$\sum_{i=1}^{M} \sum_{m \in S_i} \left\| \boldsymbol{f}_m - \mu_{q_i} \right\|^2, \tag{2.13}$$

where μ_i represents the geometric centroid of training descriptors in $\boldsymbol{S}_i$, and q_i is the data to mean assignment. By using the idea of clustering, the sum of squares of distances between the descriptors and the associated cluster centroid can be minimized. There are two main algorithms proposed for k-means clustering. An optimization technique based on Lloyd's algorithm is the first one that computes the best possible means μ_k using given assignments q_i, and computes the best possible assignments q_i given the means μ_k, i.e., $q_{ki} = \arg\min_k \|\boldsymbol{f}_m - \mu_k\|^2$. The second approach is based on an approximated version of Lloyd's algorithm [116], where the best assignments are made by using the nearest-neighbor algorithm. This approach is utilized for large size vocabularies.

2.7.2 Gaussian Mixture Model (GMM) based clustering

Gaussian mixture model [117] is based on probability density function which assumes that all the data samples can be represented in terms of mixture of Gaussian functions with unknown parameters. It contains some useful information about the covariance matrix of the data and latent Gaussian function. It is also considered as the generalization of *k*-means algorithm. GMM uses expectation-maximization (EM) algorithm to compute the unknown parameters including the prior probability value π_k, covariance matrices $\Sigma_k \in \mathbb{R}^{D \times D}$, and the mean $\mu_k \in \mathbb{R}^D$. The probability density function is defined as follows:

$$p(\boldsymbol{f}|\theta) = \sum_{k=1}^{K} p(\boldsymbol{f}|\mu_k, \Sigma_k)\pi_k \ , \tag{2.14}$$

$$p(\boldsymbol{f}|\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^D det\Sigma_k}} e^{-\frac{1}{2}(\boldsymbol{f}-\mu_k)^T \Sigma_k{}^{-1} (\boldsymbol{f}-\mu_k)} . \tag{2.15}$$

Data to cluster assignments by GMM is defined as follows:

$$q_{ki} = \frac{p(\boldsymbol{f}_i|\mu_k, \Sigma_k)\pi_k}{\sum_{j=1}^{K} p(\boldsymbol{f}_i|\mu_j, \Sigma_j)\pi_j} . \tag{2.16}$$

2.7.3 Feature Encoding Techniques

Bag of features (BOF) [112] was one of the earliest proposed feature-encoding method, which provides superior performance in solving image classification problems. It consists of three major steps: (1) Feature detection; (2) Feature representation, and (3) Codebook learning. BOF can be defined as the histogram representation of local features. In local features, SIFT has been proved as a state-of-the-art descriptor, which is invariant to rotation, translation, and transformations of the data samples. An image is first divided into a number of patches, which are then represented by 128-dimensional feature vector. The third and the last step of BOF model is to convert extracted local features into visual codewords, which is performed by learning a codebook. As we discussed before, K-means clustering [115] is considered as one of the most popular technique to initialize a codebook, which is done by dividing feature vectors into various clusters. The centers of the learned clusters represent the visual codewords. The modeling capacity can be improved by using a large size codebook, which is actually the number of clusters. However, it does not consider the layout structure of the features, which makes it unsuitable for capturing the shape of an object.

To solve this problem, many extensions of BOF model [113, 118, 119] were proposed, which can be categorized into two categories. These two categories are generative models and discriminative models, respectively. These methods have achieved superior performances by using spatial pyramid matching (SPM) [113]. This matching approach first computes feature descriptors from densely located feature points, and then applies the learned vocabularies or codebook with N entries to convert the descriptors into an N-dimensional codeword. To further increase the scalability, Yang et al. [120] proposed using sparse coding to obtain non-linear codes for non-linear feature representations. Yu et al. [121] achieved an improvement to the sparse coding (SC)-based approach by proposing a model, namely local coordinate coding (LCC), which performs feature-encoding based on the locality information. Similar to SC, the model also needs to solve the l_1-norm minimization problem, which makes it computationally expensive. A fast implementation of LCC was proposed in [122], which incorporates the locality information while encoding. The method search for n number of nearest neighbors (minimum distances) between the query image's features and the learned codebook, which accelerates the process of feature encoding. Furthermore, it also preserves the local structural information, which favors both better feature representation and classification.

2.7.3.1 Locality constrained linear coding

Locality constrained linear coding (LLC) [122] is an efficient coding technique, which projects feature descriptors to the local linear subspace using locality constraint. The final feature representation is obtained by performing max-pooling operation on the projected features. The method claimed that sparsity can be achieved using locality information, but not vice versa. Suppose $\boldsymbol{F}$ is a set of local feature descriptors, i.e. $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_M] \in \mathbb{R}^{D \times M}$. By learning a codebook having N entries, $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_N] \in \mathbb{R}^{D \times N}$, it converts extracted local features into a N-dimensional codeword for better image representation. To perform feature encoding, the following objective function is minimized:

$$\min_{\boldsymbol{W}} \sum_{i=1}^{M} \|\boldsymbol{f}_i - \boldsymbol{W}\boldsymbol{c}_i\|^2 + \lambda \|\boldsymbol{l}_i \Theta \boldsymbol{c}_i\|^2, \tag{2.17}$$

where Θ is an element-wise multiplication operator, $\boldsymbol{l}_i \in \mathbb{R}^N$ is an exponential locality adaptor, and can be defined as follows:

$$\boldsymbol{l}_i = exp\left(\frac{dist(\boldsymbol{f}_i, \boldsymbol{W})}{\sigma}\right), \tag{2.18}$$

where $dist(\boldsymbol{f}_i, \boldsymbol{W})$ is the Euclidean distance between each local descriptor $\boldsymbol{f}_i$ and codebook $\boldsymbol{W}$. σ is a constant value, that controls the weight decay speed of the locality adaptor. To speed-up the encoding process, LLC utilizes the *k*-NN search strategy by selecting specific number of nearest neighbors as a local bases from the codebook. The approximated LLC solution can be written as:

$$\min_{\widetilde{\boldsymbol{W}}} \sum_{i=1}^{M} \|\boldsymbol{f}_i - \widetilde{\boldsymbol{c}}_i \boldsymbol{W}_i\|^2. \tag{2.19}$$

Equation (2.19) can be rewritten as:

$$arg \min_{\boldsymbol{C}, \boldsymbol{W}} \sum_{i=1}^{M} \|\boldsymbol{f}_i - \boldsymbol{W}\boldsymbol{c}_i\|^2 + \lambda \|\boldsymbol{l}_i \Theta \boldsymbol{c}_i\|^2. \tag{2.20}$$

The major objective is to represent each local feature descriptor as a product of a codebook and an LLC code. Firstly, codebook is initialized using k-means clustering, and then coordinate descent method is utilized to iteratively optimize the value of $\boldsymbol{C}$ based on the existing value of $\boldsymbol{W}$, and vice versa. It is proved to be computationally efficient as compared to sparse-coding-based approach, as it does not need to solve any l_1 minimization problem for encoding.

2.7.3.2 Fisher Vector encoding

The roots of fisher vector (FV) [123] can be derived from fisher kernel, which is actually used to perform comparative analysis of two samples produced by a generative model. Till now, various extensions [124, 125] of FV have been proposed. All these methods utilize GMM as the generative model of local features. It has been proved to be effective in modeling of low-dimensional features, such as SIFT. The number of Gaussian mixtures needed for feature modeling depends on the volume of feature space. Recently, high-dimensional features, such as deep neural networks [126], high-dimensional LBP [127], pooled feature vectors [128, 129], etc. gained a lot of attention due to their superior performance. To model the high-dimensional feature space, a large number of Gaussian mixtures are needed. Now, we will review the general formulation of FV coding along with its limitations.

Let us assume that we have two samples generated using a generative model, Fisher kernel can be utilized to compute the similarity between them. Instead of using data matrices directly, set of local features are extracted from data samples, denoted as $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_N\}$. Each feature can be modeled by its corresponding probability density function. According to fisher kernel, a feature $\boldsymbol{F}$ can be represented by its gradient vector computed over the model parameter λ.

$$\boldsymbol{G^F}_\lambda = \nabla_\lambda \log P(\boldsymbol{F}|\lambda) = \sum_j \nabla_\lambda \log P\left(\boldsymbol{f}_j\middle|\lambda\right) . \tag{2.21}$$

The fisher kernel can be written as $K(\boldsymbol{F}, \boldsymbol{Y}) = \boldsymbol{G^F}_\lambda{}^T \boldsymbol{I}^{-1} \boldsymbol{G^F}_\lambda$, where $\boldsymbol{I}$ is an information matrix, defined as $\boldsymbol{I} = E[\boldsymbol{G^F}_\lambda \boldsymbol{G^F}_\lambda{}^T]$. Information matrix $\boldsymbol{I}$ can be excluded to reduce the computational complexity. To compare two feature vectors using fisher vector encoding (FVC) method, first step is to compute their corresponding gradients, followed by sum-pooling. Therefore, the resultant feature vector can be considered as an encoded feature. However, there are some limitations associated to the FVC. To implement this framework, there is a need to define the distribution $P(\boldsymbol{f}|\lambda)$. As we discussed above, most of the methods utilizes GMM as a generative model for feature $\boldsymbol{f}$. First, a Gaussian model $\mathrm{N}(\mu_k, \Sigma_k)$ is drawn from the prior distribution $P(k)$, where $k = 1,2, \dots, n$. After that, the extracted local feature $\boldsymbol{f}$ is drawn from $\mathrm{N}(\mu_k, \Sigma_k)$. In general, the feature $\boldsymbol{f}$ within a local region of a feature space follows a Gaussian distribution. It means that each Gaussian mixture only covers a little portion in the whole feature space. For low-dimensional features, such as SIFT, only few hundreds of Gaussian mixtures are needed. For high-dimensional features, this number is insufficient, which ultimately results in inaccurate modeling. The method in [130] proposed a sparse-coding based

FVC method, which utilizes infinite number of Gaussian mixtures. It is based on the assumption that each local feature follows a Gaussian distribution, and has a random mean vector. This mean vector can be considered as a point on a subspace, that can be defined by a set of bases of an over-complete dictionary, and it is indexed by a coding vector $\boldsymbol{u}$. The coding vector $\boldsymbol{u}$ is drawn from a Laplacian distribution function $P(\boldsymbol{u}) = \frac{1}{2\lambda}\exp\left(-\frac{|\boldsymbol{u}|}{\lambda}\right)$. After that, local feature is drawn from a Gaussian distribution $\mathrm{N}(\boldsymbol{Bu}, \Sigma)$. Here the Laplacian prior ensures the fisher vector to be sparse. After establishing the generative model of local features, the fisher coding vector can be derived by taking the derivative of its log-likelihood, which can be written as:

$$C(\boldsymbol{f}) = \frac{\partial \frac{1}{\sigma^2}\|\boldsymbol{f} - \boldsymbol{B}\boldsymbol{u}^*\|_2^2 + \lambda\|\boldsymbol{u}\|_1}{\partial \boldsymbol{B}}, \tag{2.22}$$

where $\boldsymbol{u}^* = \arg\max_{\boldsymbol{u}} P(\boldsymbol{f}|\boldsymbol{u}, \boldsymbol{B})P(\boldsymbol{u})$. The further mathematical derivation can be found in [130]. After getting the encoded features, pooling and normalization operations are applied. Most of the feature encoding methods utilizes handcrafted feature descriptors, such as SIFT. However, the utilization of high-dimensional deep features can bring a significant improvement in recognition accuracy, as reported in [130].

2.8 Face Recognition using sparse representation

Face recognition can be considered as a complex high-dimensional pattern recognition problem. Sparse-coding aims to represent an image by selecting least number of atoms from an over-complete dictionary. Given the number of training images from the j^{th} class, $\boldsymbol{A}_j = [\boldsymbol{u}_{j,1}, \boldsymbol{u}_{j,1}, \dots, \boldsymbol{u}_{j,n_j}]$, any test image $\boldsymbol{y}$ from the same class can be linearly represented in terms of training samples from the same class, i.e. $\boldsymbol{y} = \alpha_{j,1}\boldsymbol{u}_{j,1} + \alpha_{j,2}\boldsymbol{u}_{j,2} + , \dots + \alpha_{j,n_j}\boldsymbol{u}_{j,n_j}$. In face recognition, we don't know the class of the testing sample, so a new matrix $\boldsymbol{A}$ is defined that contains n training samples of all k classes, i.e. $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \dots, \boldsymbol{A}_k] = [\boldsymbol{u}_{1,1}, \boldsymbol{u}_{1,2}, \dots, \boldsymbol{u}_{k,n_k}]$. So, now a test image $\boldsymbol{y}$ can be linearly represented in terms of the whole training set as $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0$, where $\boldsymbol{x}_0 = [0, \dots, 0, \alpha_{j,1}, \alpha_{j,2}, \dots, \alpha_{j,n_j}]$ is an associated sparse coefficient vector, which has non-zero values only for the samples associated with the j^{th} class, while zero for the other remaining classes. The associated sparse coefficients are determined using l_0 norm. As l_0 norm is an NP-hard problem, so l_1 norm is used to compute these sparse coefficients. Mathematically, it can be written as:

$$\min_{\alpha}\|\alpha\|_1 \quad \text{s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \varepsilon, \tag{2.23}$$

where $\boldsymbol{y}$ is a query image, $\boldsymbol{A}$ is a dictionary, and α is the corresponding sparse coefficient vector. In case of noise variations, sparse coefficients with non-zero values can be obtained for more than one subjects in the training set. To solve this issue, one can classify $\boldsymbol{y}$ to the class having a largest non-zero value. However, this is not a reliable solution. Therefore, a test sample $\boldsymbol{y}$ is classified based on how well the obtained sparse coefficients can reconstruct $\boldsymbol{y}$ with respect to the all training samples. So, it determines the similarity between the query image and the training samples from each class by computing the residual values. A query image is assigned to the class which has least residual value (reconstruction error). SRC proves to be quite efficient in recognizing face images suffered by noise, occlusion or disguise. In terms of face recognition, SRC focuses on two major issues; (1) Feature selection, and (2) robustness to noise, disguise and occlusion. Some more work [131, 132] based on sparse coding was proposed to tackle the pose and lighting variations in face recognition. To further increase the robustness to occlusion, Yang et al. [133] proposed to use Gabor features instead of original data matrices, and then utilizes sparse coding to solve occluded face recognition problem, which brings significant improvement in recognition accuracy. Yang et al. [134] proposed another sparse coding-based algorithm, which argues that sparse coding can be considered as a robust regression problem. The method assumes that the coding residuals follows either Laplacian or Gaussian distribution, which is not feasible for removing coding errors. Therefore, RSC [134] looks to estimate the maximum likelihood solution of sparse coding problem, which is proved to be more robust to noise, disguise, and occlusion. In addition to face recognition, sparse-coding has also provided superior performance in other machine learning tasks, such as image super-resolution [62], object detection, etc., which makes it a hot research topic for researchers around the world.

2.9 Low-rank Matrix decomposition and its applications to Face Recognition

In real world environments, the data available for training and testing may contain some random noisy components. Previously proposed methods, such as PCA [1], LDA [26], and SRC [7] assumes that the training data is captured under controlled environment, and do not contain any kind of contamination. This degrades their recognition performance when corrupted testing data is presented for recognition. Recently, methods based on robust PCA [135-137] were proposed, which deals with the images containing random noise. Among all these methods, the low-rank approximation techniques have proven to be highly robust against various noise variations. It works by decomposing the corrupted data matrix into a clean low-rank matrix and a sparse error matrix. As we discussed earlier, SRC classifies a test sample by computing a minimum reconstruction error with respect to a training set. It is observed that if the

training set contains noise or occlusion, the performance of SRC degrades heavily. In addition to SRC, Wagner et al. [131] proposed to solve the face misalignment problem, using sequential l_1 minimization algorithm. The method also handles lighting variations by using a projector-based illumination system. Furthermore, Zhou et al. [138] solves occluded face recognition problem by integrating Markov random field with SRC. However, none of these methods considers the possible corruption in training data, which makes them unsuitable for read-world applications. To solve this problem, robust face recognition method was proposed in [139, 140], which learns PCA subspace using recovered low-rank part for classification. However, face images of different subjects might share some similarities in their features, e.g. location of landmark points (eyes, nose), etc. Therefore, the learned low-rank part might not be enough to discriminate between the identities of the two face images. Therefore, Wei et al. [140] regularized the objective function with structural incoherence constraint. The constraint suppressed the shared features of the two different subjects, while preserving the discriminative ones. The method can be more effective, if some discriminative handcrafted or deep features are extracted prior to perform matrix decomposition. Jing et al. [141] proposed to solve multi-spectral face recognition problem when face images are contaminated by noise. The method learns a multi-spectrum low-rank dictionary, which explores the correlation as well as complementary information between the different spectrums. Wu et al. [142] proposed an image classification method, which learns a multi-view low-rank dictionary to remove noise components in multiple views. However, the global & local structural properties of the data samples are not considered, while projecting the recovered low-rank data onto a low-dimensional subspace. As discussed earlier, face images often lie in a non-linear manifold space. By learning this manifold space, images can be classified with high precision. Therefore, there is a need to incorporate the manifold information, which considers the geometrical structure of the data samples, while learning the low-rank features.

2.10 Analysis of Periocular Regions for face recognition

The periocular region of a human face contains complex biomedical features, such as eyebrows, contour, eyeballs, eyelids, etc. From the biological point of view, the high complexity of any region leads to more coding processing, which means that the appearance contains more protein and genes and more discriminative information. This encourages researchers to consider the periocular region as the most discriminative region on a human face to differentiate among different people. In this regard, researchers have proposed various techniques to investigate the discriminative power of the periocular region to solve many biometric problems. Xu et al. [143] extracted local feature

descriptors from the periocular region of a human face for solving the age-invariant face recognition problem, and achieved a very high recognition rate on one of the most challenging face-aging datasets, FGNET [144]. It argues that a periocular region undergoes a very little effect over time as the shape and location of the eyes remains the same, while cheeks, chin, nose, etc. go through significant changes with age progression. Methods proposed in [145, 146] performs gender classification using periocular region. Most of the existing work employs local feature descriptors, e.g. Gabor wavelets, LBP, HOG, SIFT, etc. and then used SVM for classification. The method proposed in [147] presented a face recognition system, which extracts Gabor features from all of the facial landmarks and utilizes a different classifier for each landmark. The final recognition result is obtained by fusing the outputs of all the classifiers. Miller et al. [148] proposed to extract the LBP features from both the periocular region and the whole face. The method demonstrated that if the image quality is extremely poor, then the periocular region could provide better performance than the whole face region. Park et al. [149] applied masking below the nose to study the partial occlusion problem. The results revealed that the recognition performance heavily degrades due to the occlusion. However, the periocular region proves to be more robust against occlusion than any other facial region. Researchers in [150] matched face images, captured before and after plastic surgery by combining the scores of the complete face and the periocular region, and obtains the rank-1 recognition accuracy of 87.4%. Furthermore, the method in [151] extracts features from different face regions to study the facial variations due to gender transformation. The results show that the periocular region outperforms other regions in terms of discriminative power and obtains the highest recognition rate. However, the performance of these handcrafted features depends on proper preprocessing operations, including contrast enhancement, pose correction, illumination normalization, etc. In most of the methods, the region of interest (ROI) is defined around the eye region to extract the corresponding features. However, some components in this region are not relevant to recognize the identity, such as hairs, and glasses. The reason for this is that the extracted features are not enough discriminative across all the regions. In one of our proposed frameworks, we investigate the periocular region of a human face to perform age-invariant face recognition, which will be explained later in this thesis.

2.11 Conclusions

This chapter presents the brief overview of the existing work in face recognition research. Some state-of-the-art techniques are reviewed including deep-learning based approaches, feature-encoding methods, subspace learning techniques, sparse representation, and low-rank matrix decomposition. Furthermore, the discriminative power of the

periocular region of a human face was briefly discussed. Particularly, existing literature on low-resolution, and aging face recognition are extensively studied. In the coming chapters, we will briefly discuss our proposed solutions for solving the problems discussed in this chapter.

Chapter 3 Learning Sparse Discriminant Low-rank Features for Low-resolution Face Recognition

3.1 Introduction

Low-resolution (LR) face recognition is a challenging research problem in the field of machine learning. It has a huge demand in various surveillance applications. Although, remarkable progress has been made in recognizing face images captured under constrained environments, the problem of identifying the poor-quality images taken by security cameras is still unsolved. As we discussed in Chapter 2, performances of face recognition methods heavily depend on the amount of discriminant, robust features that can be extracted from face images, which reside in high-frequency components. However, when the image resolution decreases, the information available for distinguishing faces becomes less, and therefore Super-resolution (SR) becomes ineffective under unconstrained variations. This is because reconstruction from low-frequency components is an ill-posed problem and creates artifacts in the super-resolved images. Performances of the recognition-based and feature-based super-resolution methods depend on the feature extraction technique and reconstruction regularization model being used, but it is still unclear which regularization methods are optimal from the recognition perspectives. In addition to that, the choice of features, which need to be able to handle large amount of variations in unconstrained environments, is critical. As we reviewed in the previous chapter, most of the methods based on coupled mappings operate directly on data matrices and do not extract the robust features from face images while projecting the data samples into a unified feature subspace. This reduces the recognition accuracy in uncontrolled settings, especially under pose and illumination variations. Robustness of face recognition systems (FRS) can be greatly improved by using a combination of local features. Now, the question is which features should be selected and combined? In this chapter, we address LR face recognition problem by fusing two local feature descriptors, which are robust to various facial variations. As we discussed in the previous chapter, low-rank matrix has better feature representation ability as compared to original data matrix, so we utilize the low-rank matrix-decomposition algorithm [137] to convert the extracted fused features into a low-rank matrix and a corresponding error matrix (sparse in nature). For recognition, we only utilize the low-rank component, while discarding the sparse error matrix. Furthermore, our approach first down-samples gallery faces to the size of the query image, and then perform recognition. Although the super-resolved face image or feature contains more information, the estimated information may be incorrect and distorted. By down-sampling the gallery faces, no prediction is necessary. In this chapter, we

propose an effective solution for tackling all the variations simultaneously, by utilizing the discriminative power of sparse representation of multiple low-rank local features for LR face recognition.

This chapter is structured as follows. Section 3.2 discusses the motivation behind our proposed approach. Section 3.3 introduces our proposed morphological pre-processing method, then Gabor wavelets and LBPD features are described. Section 3.4 explains the concept of low-rank feature learning. Section 3.5 presents our proposed framework based on sparse coding. Section 3.6 introduces the linear regression model used for classification. Section 3.7 provides brief description of our experimental setup and results. Finally, we conclude our chapter in Section 3.8.

3.2 Motivation behind the proposed idea

Motivated by the applications of sparse representation [7, 152] to pattern recognition, we propose a new approach for solving the LR face recognition problem based on sparse coding of multiple low-rank local features. The major idea is to compute an optimum sparse matrix, which projects the gallery and query low-rank features onto a common low-dimensional subspace for recognition. Sparse coding provides natural discriminant power and represents face images in a compact manner. There exists a linear relationship between a test sample and the other training samples of the same subject. Matching a HR gallery image with a LR probe image has the dimension-mismatch problem, which also produces noise while learning a unified feature subspace. Our proposed method first down-samples all HR gallery images to the same resolution as the LR probe (test) image, and performs recognition in the LR domain. In our proposed approach, we assume that two images of the same resolution have higher correlation as compared to having two different resolutions. There are two reasons for this. First, low-dimensional features are effective in computing the within-class as well as between-class scattering matrices, as their dimension is lower than, or closer to, the total number of samples available. Second, low-frequency spectrum carries the information regarding the illumination variations, so lighting conditions can be improved by utilizing the low-frequency information. In [9], it has also been argued that down-sampling both the training and testing images can increase the recognition rate, even for images of very low resolution, such as 6×6 pixels. This proves that down-sampling face images is a feasible approach to solve the dimension-mismatch problem.

As discussed in the previous chapter, SR algorithms are not feasible for recognition purposes. This is because of the generation of artifacts in the super-resolved images, which reduce the recognition accuracy. To overcome this problem, a new morphological pre-processing approach based on top and bottom-hat filtering is proposed, which

improves an image quality, without generating any kind of distortion or artifacts in the final processed image. To make our approach robust to variations in unconstrained environments, two local features, Gabor wavelets and Local Binary Pattern Difference (LBPD) [153], are extracted from both the training and testing face images followed by normalization. After that, we perform feature-level fusion to form a final normalized feature vector. The normalized fused features are used to learn a low-rank matrix and a sparse error matrix, using an augmented Lagrangian method. The extracted low-rank matrix is then utilized to learn a new low-dimensional feature subspace by computing a projection matrix based on our proposed sparse-coding-based algorithm, such that the sparsity of the learned features is preserved. After that, the similarity between the gallery and query features is determined by estimating a coefficient vector using linear regression [43]. Based on the coefficient vector, residuals are computed for feature matching. To increase the discriminability between the face images of two different subjects, class-label information is utilized. Furthermore, our method has less computational complexity than other linear and nonlinear mapping-based methods [154, 28]. Our method can estimate local structures of face images by utilizing the sparse prior knowledge. Extraction and fusion of multiple low-rank local features make our method effective for recognition in unconstrained environments.

3.3 Proposed Framework for LR Face Recognition

3.3.1 Pre-Processing and Feature Selection

Face alignment and normalization are considered as the most important steps prior to face recognition. Furthermore, face images are usually pre-processed so that they can be more standardized. This can help improve the recognition rate. Now, we will describe our proposed morphological pre-processing method, and the features selected for our proposed algorithm.

3.3.1.1 Morphological pre-processing

To solve the problem of artifacts and distortion produced by SR algorithms, a novel morphological pre-processing method based on top and bottom-hat filtering is proposed. This method not only extracts useful information from face images, but also eliminates low-contrast features. It can also alleviate the effects of non-uniform illumination, so it is suitable for tackling variations in lighting conditions.

Let I be a grayscale image, and s be a disk-shaped structuring element with radius R. The first step is to apply the top-hat filtering which is defined as $T_t(I) = I - I \circ s$, where $\circ$ represents the opening operation. It can extract bright features from an image. Similarly, dark features are extracted by bottom-hat filtering defined as $T_b(I) = I \bullet s - I$, where $\bullet$ represents the closing operation. To enhance the local contrast for better image understanding, the face image is added to the difference between the two filtering outputs. Mathematically, it is defined as follows:

$$I_{CE} = I + T_t(I) - T_b(I), \tag{3.1}$$

where $T_t(I)$ and $T_b(I)$ represent the top and bottom-hat filtering images, respectively, and I_{CE} is the contrast-enhanced face image.

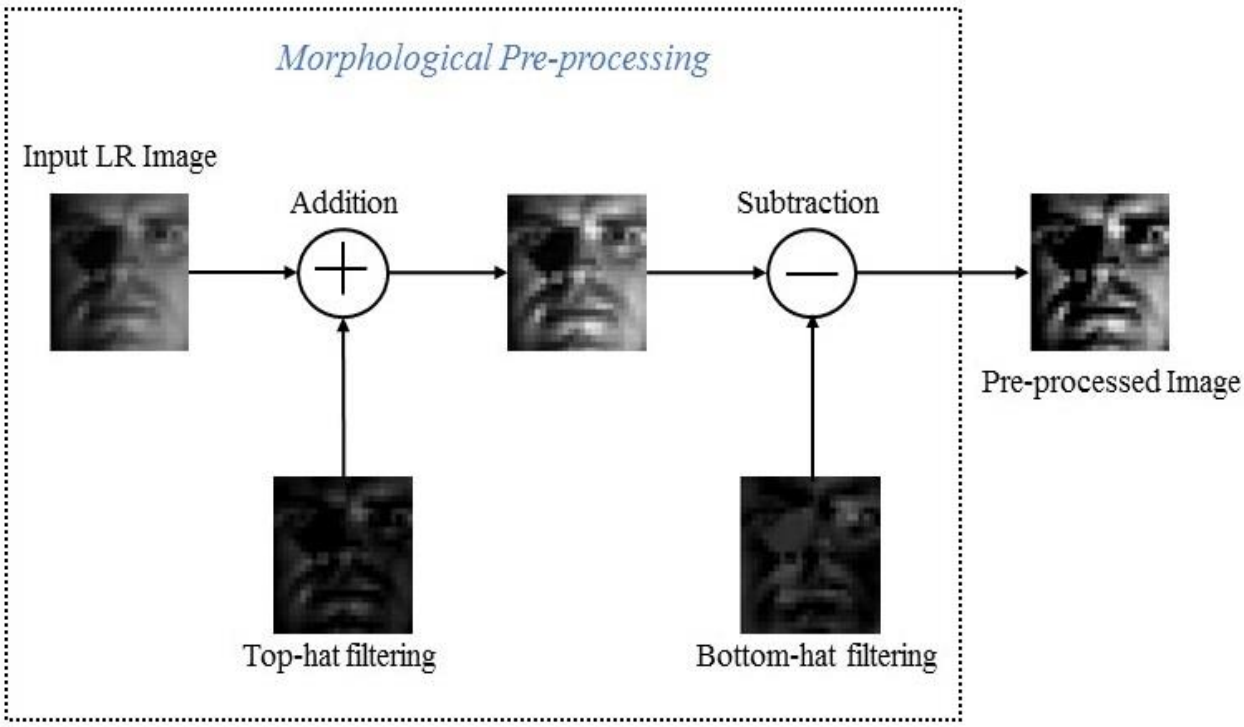
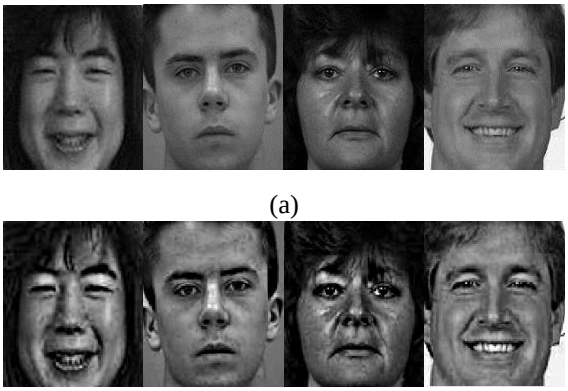


Fig. 3-1. The morphological pre-processing steps, on a LR face image.

Fig. 3-1 shows a LR face image with uneven illumination and the corresponding contrast-enhanced image generated by the proposed filtering operation. Fig. 3-2 shows HR face images, and the images obtained after applying our proposed morphological operation. It can be seen that our method provides face image with better visual quality, which can be helpful in extracting more useful facial information for recognition.



(a)

(b)

Fig. 3-2. (a) Sample HR face images, and (b) Pre-processed images using proposed morphological pre-processing scheme.

3.3.1.2 Gabor Wavelets

From the biological point of view, Gabor functions can be used to model the responses of the cells in the visual cortex of mammalian brains. Furthermore, it exploits the local regions to extract information at multiple scales and orientations. It is an efficient local feature descriptor, which has been proved to be quite useful in various computer vision applications, such as object tracking, detection, and recognition. Gabor wavelets (GWs) can be used for spatial-frequency analysis because they have both the multi-orientation and multi-resolution properties, which enables it to provide valuable information about the local structure of an image. GWs achieve optimal representation in the frequency domains. It is defined as a complex exponential modulated by a Gaussian function, which is written as follows:

$$\phi_{\omega,\theta}(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)} \cdot \left[e^{i(\omega x cos\theta + \omega y sin\theta)} - e^{-\frac{\omega^2\sigma^2}{2}} \right], \tag{3.2}$$

where (x, y) represents the pixel positions, ω is the frequency of the sinusoidal plane wave, θ represents the orientation, and σ is the standard deviation corresponding to the Gaussian envelope. In our algorithm, we extract features at five scales and eight orientations.

3.3.1.3 Local Binary Pattern Difference Feature

Local binary pattern (LBP) feature [155] is extracted by first partitioning an image into a number of blocks. In each block, the LBP code at each pixel position is generated, by comparing the central pixel with its corresponding neighboring pixels residing on a circle of radius R, centered at the pixel under consideration. If a neighboring pixel has its value smaller than the central one, then it is labeled as '0', otherwise as '1'. A string of the binary bits is used to

form an LBP code for the pixel. Mathematically, the LBP code of the central pixel x_c with respect to the channel ϕ can be defined as follows:

$$I_{LBP_{N,R}}(x_c, \phi) = \sum_{n=0}^{N-1} u\big(\phi(x_n) - \phi(x_c)\big)2^n, \tag{3.3}$$

where x_n $(n = 0, \dots, N-1)$ represents the N neighboring pixels on the circle of radius R centered at pixel x_c; ϕ can be either intensity value or filter response of an image; and $u(x)$ is the step function, i.e. its value is '1' if $x \geq 0$, and '0' otherwise. There are many LBP variants [156-158], but LBP cannot combine with other features for recognition. The problem with LBP is that the LBP code is a non-numerical representation, which is a discrete pattern rather than a numerical response. Recently, a numerical variant of LBP [153], which is known as local binary pattern difference (LBPD), was proposed. To extract this feature, the mean LBP of a given region is computed, then the LBPD at a pixel position is computed as the difference between its LBP code and the mean LBP. The Karcher mean [159] is used to compute the mean LBP of a region, which minimizes the sum of distances to all the points in a given image region. Each element of a binary vector $\hat{\boldsymbol{I}}_{LBP}$ represents a specific bit of the regular LBP. Specifically, the k^{th} bit of $\hat{\boldsymbol{I}}_{LBP}$ is given as follows:

$$\hat{\boldsymbol{I}}_{LBP}(k) = u\big(\phi(x_k) - \phi(x_c)\big). \tag{3.4}$$

Suppose that there are P LBPs in a region, represented as $L = \{\hat{l}_1, \hat{l}_2, \dots, \hat{l}_P\}$. The k^{th} element for $k = 0, \dots, K-1$ of its Karcher mean $\widehat{\boldsymbol{m}}_I$ is defined as follows:

$$\widehat{\boldsymbol{m}}_I(k) = \left\lfloor \frac{\sum_{p=1}^{P} \hat{\boldsymbol{I}}_p(k)}{P} + 0.5 \right\rfloor, \tag{3.5}$$

where $\lfloor . \rfloor$ is the floor function, and $\widehat{\boldsymbol{m}}_I$ belongs to the set of the 2^K LBPs. To relax the constraint that the LBP mean is an LBP, the mean LBP vector can be a floating-point vector denoted by $\widehat{\boldsymbol{m}}_f$, as follows:

$$\widehat{\boldsymbol{m}}_f = \frac{\sum_{p=1}^{P} \hat{\boldsymbol{I}}_p}{P}. \tag{3.6}$$

(a) LBP Difference

Let us consider the LBP code $\hat{\boldsymbol{I}}$ and mean $\widehat{\boldsymbol{m}}_f$ of a face image. The LBPD feature vector can be computed as $\widehat{\boldsymbol{d}} = \hat{\boldsymbol{I}} - \widehat{\boldsymbol{m}}_f$. Magnitude of the LBPD feature is given by:

$$I_{LBPD}{}^{s}(x,\phi) = \left\|\hat{\boldsymbol{I}} - \hat{\boldsymbol{m}}_f\right\|, \tag{3.7}$$

where $\|.\|$ can be of any type of norms. Its values are positive, so it is also known as unsigned LBPD. To extract more discriminative information from an image, the sign is introduced by defining the LBPD feature as follows:

$$I_{LBPD}{}^{s}(x,\phi) = s(\left\|\hat{\boldsymbol{I}}\right\| - \left\|\hat{\boldsymbol{m}}_f\right\|)\left\|\hat{\boldsymbol{I}} - \hat{\boldsymbol{m}}_f\right\|, \tag{3.8}$$

where $s(x)$ represents the signum function, whose value is 1 if $s(x) \geq 0$ and -1 if $s(x) < 0$. This will form an ordered LBP feature vector. It is not affected by the permutation of bits, which makes it rotation invariant. Fig. 3-3 show two face images from the LFW database and their corresponding LBPD images and histograms.

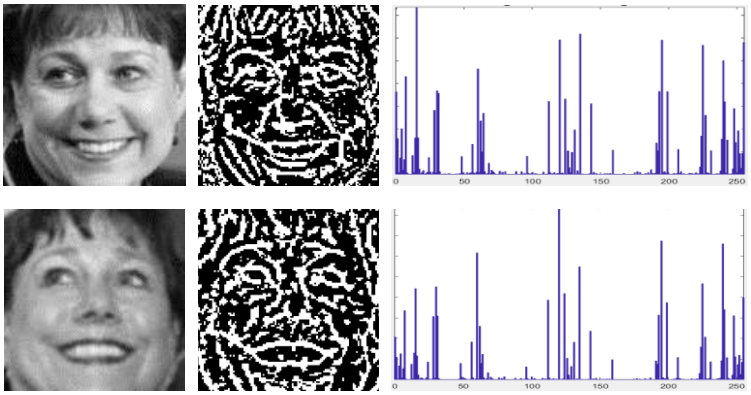


Fig. 3-3. LBPD feature and histogram representation of two face images.

3.3.1.4 Features Selection

Selection of appropriate feature descriptor is quite important to achieve optimum performance in image classification and object recognition. In the last two decades, various global and local feature descriptors have been proposed. As discussed earlier, local features tend to outperform global features, and have been proven to be more robust against various geometric variations. Some of the state-of-the-art feature descriptors include SIFT [160], SURF [161], HOG [162], Gabor [163], and LBP [155]. To improve the performance, various extensions of these descriptors have been proposed. For facial image analysis, Gabor and LBP have been proved to be the best performing feature descriptors for face recognition [164]. For object recognition, the most widely used local feature descriptor is SIFT, which first extracts relevant keypoints from given images, and then represents the gradient information in the neighborhood of each keypoint. This feature exhibits both scale and rotation invariance. However, its major drawback is high computational complexity. Inspired by SIFT, a faster version known as SURF was proposed, whose performance strongly depends on the relative keypoints that can have a variable geometry. It is highly desirable that the selected

feature descriptor has high discriminative power and low computational complexity. In comparison to SIFT, LBP is simple and fast to compute. It can efficiently describe the local texture information, while showing high robustness to monotonic gray-level transformations. Moreover, the features computed by using LBP are fixed relative to each other and can better distinguish between the curved surfaces, such as face images. For a difficult task, such as LR face recognition, using a single feature is unable to capture sufficient discriminative information from face images. In [164], it was argued that combining the LBP with Gabor features can enhance the recognition performance, up to a significant level. As discussed in the previous section, the LBP feature consists of discrete patterns or symbols, rather than a numerical response, so the LBP feature cannot combine with other features directly.

In our proposed method, we employ two efficient texture descriptors (Gabor wavelets and LBPD), due to their supplementary natures, which makes them promising candidates for fusion. There are various reasons for this. First, the Gabor features can encode the facial shape information at multiple scales and orientations. Each GW may be viewed as a bandpass filter, which extracts features at a specific range of frequencies and orientations in the frequency domain. Okajima et al. [165] argued that Gabor wavelets can be used as a solution for the mutual information maximization problem. By using Gabor-type receptive field, maximum amount of information can be extracted from local regions. It is clear that image rotation affects the permutation of bits. According to Equation 3.3, LBP used the predetermined weights to weigh the bits, which results in the different LBP codes of the original image and its rotated version. Therefore, extra effort is necessary to achieve rotation invariance. LBPD is inherently rotation-invariant, as the norm employed in Equations (3.7) and (3.8) makes sure that the code does not depend on permutation of the bits. Furthermore, LBPD consists of numerical responses, whereas LBP is a collection of discrete patterns. This numerical property of LBPD makes it attractive in terms of texture analysis. LBPD does not consider the intensity of pixels, because it utilizes the sign of comparisons between the neighboring pixels, as in Equation (3.4). This makes LBPD invariant to lighting conditions. Due to the abovementioned properties of Gabor and LBPD features, better feature representation can be obtained, which is invariant to various facial variations.

3.4 Low-rank Feature learning

Recently, low-rank matrix recovery has gained plenty of attention due to its number of applications in many machine-learning tasks, such as face recognition [139, 140], data mining, image classification [166], etc. Instead of using data matrices directly, numerous kinds of features can first be utilized to get useful information from images that

provides better representation. As discussed in [166], extracted local features may exhibit some noisy patterns, which can reduce the recognition performance. Motivated by this observation, we decompose the extracted fused feature vectors $\boldsymbol{F}$ into a low-rank feature matrix $\boldsymbol{L}$ and a corresponding sparse error matrix $\boldsymbol{S}$. The extracted low-rank feature matrix $\boldsymbol{L}$ has been proved to be more discriminative for recognition, as it provides better feature representation. It works by minimizing the rank of the matrix $\boldsymbol{L}$, while computing the l_0-norm of $\boldsymbol{S}$. It can be written as follows:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \operatorname{rank}(\boldsymbol{L}) + \lambda\|\boldsymbol{S}\|_0 \ \text{ s.t. } \boldsymbol{F} = \boldsymbol{L} + \boldsymbol{S}. \tag{3.9}$$

The second term computes the non-zero elements in $\boldsymbol{S}$. It can be simplified by replacing the first term of Equation (3.9) with the nuclear norm, and the second one with l_1-norm. The resulting objective function can be written as:

$$\min_{\boldsymbol{L},\boldsymbol{S}} \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \ \text{ s.t. } \boldsymbol{F} = \boldsymbol{L} + \boldsymbol{S}. \tag{3.10}$$

It turns out to be a convex optimization problem, with two major constraints. First, the rank of the recovered low-rank matrix $\boldsymbol{L}$ is not too large. Second, there should be a small number of non-zero elements in $\boldsymbol{S}$. In our method, we utilize Augmented Lagrange multiplier (ALM) to solve this optimization problem due to its low complexity. Let $\boldsymbol{F}$ be the fused features extracted from face images. Then the Lagrange function of Equation (3.10) is written as:

$$L_\mu(\boldsymbol{L},\boldsymbol{S},\boldsymbol{Y}) = \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 + <\boldsymbol{Y}, \boldsymbol{F} - \boldsymbol{L} - \boldsymbol{S}> + \frac{\mu}{2}\|\boldsymbol{F} - \boldsymbol{L} - \boldsymbol{S}\|_F^2, \tag{3.11}$$

where $\boldsymbol{Y}$ and μ represents a Lagrange multiplier and a penalty parameter, respectively. The matrices $\boldsymbol{L}$ and $\boldsymbol{S}$ are updated alternatively until converged, as follows:

$$\left(\boldsymbol{L}^{j+1}, \boldsymbol{S}^{j+1}\right) = \arg\min_{\boldsymbol{L},\boldsymbol{S}} L_\mu\left(\boldsymbol{L},\boldsymbol{S},\boldsymbol{Y}^j\right), \tag{3.12}$$

$$\boldsymbol{Y}^{j+1} = \boldsymbol{Y}^j + \mu(\boldsymbol{F} - \boldsymbol{L}^{j+1} - \boldsymbol{S}^{j+1}), \tag{3.13}$$

where j is the iteration index.

1) Updating $\boldsymbol{L}_i$

To update the low-rank matrix ${\boldsymbol{L}_i}^{j+1}$ of class i at the $(j+1)$st iteration, all the variables except $\boldsymbol{L}_i$ are fixed, which leads to the following equation:

$$\begin{aligned} {\boldsymbol{L}_i}^{j+1} &= \arg\min_{\boldsymbol{L}_i} L(\boldsymbol{L}_i, {\boldsymbol{S}_i}^j, {\boldsymbol{Y}_i}^j, \mu^j) \\ &= \arg\min_{\boldsymbol{L}_i} \|\boldsymbol{L}_i\|_* + <{\boldsymbol{Y}_i}^j, \boldsymbol{F}_i - \boldsymbol{L}_i - {\boldsymbol{S}_i}^j> + \frac{\mu^j}{2}\left\|\boldsymbol{F}_i - \boldsymbol{L}_i - {\boldsymbol{S}_i}^j\right\|_F^2 \end{aligned}$$

$$= \arg\min_{\boldsymbol{L}_i} \epsilon \|\boldsymbol{L}_i\|_* + \frac{1}{2}\|\boldsymbol{X}_l - \boldsymbol{L}_i\|_F^2, \tag{3.14}$$

where $\epsilon = (2\mu^j)^{-1}$ and $\boldsymbol{X}_l = 0.5(\boldsymbol{F}_i - \boldsymbol{S}_i{}^j + \frac{\mathbf{1}}{\mu^j}\boldsymbol{Y}_i{}^j)$.

According to Section 2.1 in [167], the above equation has a closed form, which is given as $\boldsymbol{L}_i{}^{j+1} = \boldsymbol{T}Z_\epsilon[\boldsymbol{R}]\boldsymbol{Q}^T$, where $\boldsymbol{TRQ}^T$ is the singular value decomposition of $\boldsymbol{X}_l$, and $Z_\epsilon[\boldsymbol{R}]$ is the elementwise thresholding of $\boldsymbol{R}$, i.e., $Z_\epsilon[\boldsymbol{R}](i,j) = z_\epsilon[\boldsymbol{R}(i,j)]$, where $z_\epsilon[r]$ is defined as

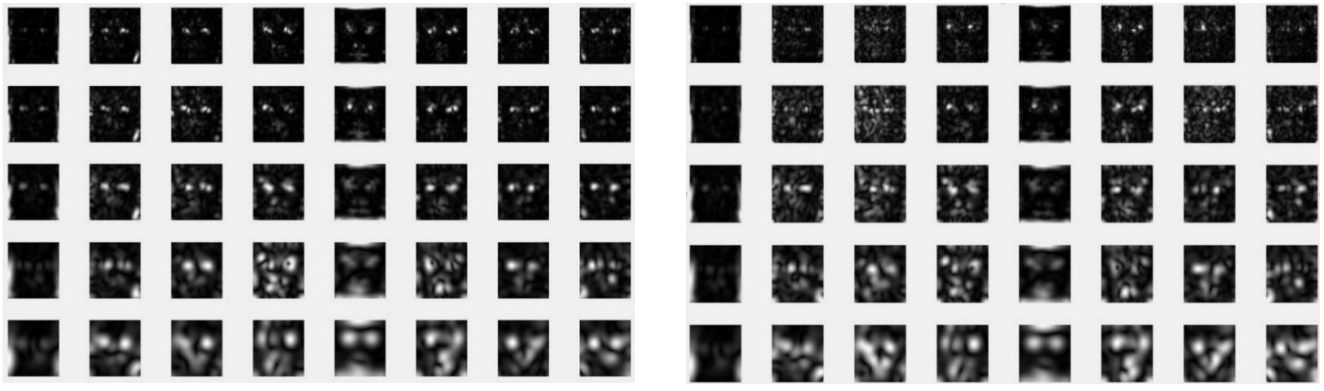
$$z_\epsilon[r] = \begin{cases} r-\epsilon, & \text{if } r > \epsilon \\ r+\epsilon, & \text{if } r < \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

2) *Updating* $\boldsymbol{S}_i$

$$\begin{aligned} \boldsymbol{S}_i{}^{j+1} &= \arg\min_{\boldsymbol{S}_i} L(\boldsymbol{L}_i{}^{j+1}, \boldsymbol{S}_i, \boldsymbol{Y}_i{}^j, \mu^j) \\ &= \arg\min_{\boldsymbol{S}_i} \lambda\|\boldsymbol{S}_i\|_1 + < \boldsymbol{Y}_i{}^j, \boldsymbol{F}_i - \boldsymbol{L}_i{}^{j+1} - \boldsymbol{S}_i > + \frac{\mu^j}{2}\left\|\boldsymbol{F}_i - \boldsymbol{L}_i{}^{j+1} - \boldsymbol{S}_i\right\|_F^2 \\ &= \arg\min_{\boldsymbol{S}_i} \epsilon' \|\boldsymbol{S}_i\|_1 + \frac{1}{2}\|\boldsymbol{X}_s - \boldsymbol{S}_i\|_F^2, \end{aligned} \tag{3.16}$$

where $\epsilon' = \left(\frac{\lambda}{\mu^j}\right)$ and $\boldsymbol{X}_s = \boldsymbol{F}_i - \boldsymbol{L}_i{}^{j+1} + \left(\frac{1}{\mu^j}\right)\boldsymbol{Y}_i{}^j$.

Similarly, the closed form solution of this optimization problem is given as $\boldsymbol{S}_i{}^{j+1} = Z_{\epsilon'}(\boldsymbol{X}_s)$. We set $\lambda = 0.001$ in our experiments. Furthermore, we discard the sparse error term $\boldsymbol{S}$, and use the low-rank approximation feature matrix $\boldsymbol{L}$ only for further processing. Fig. 3-4 shows that the recovered low-rank component of a feature can capture more facial details, as compared to the originally extracted Gabor features.



(a) (b)

Fig. 3-4. Gabor features extracted from a face image. (a) Original Gabor features with 5 scales and 8 orientations, (b) Low-rank Gabor features.

3.5 Sparse Coding of Multiple Low-Rank Features

In context of face recognition, sparse representation has gained much attention in the last decade, due to its robustness against various facial variations. Wright et al. [7] proved the effectiveness of sparse theory for recognition of face images taken in uncontrolled environments. According to the representation, a linear relationship exists between each test sample and the other training samples from the same subject, which is sparse in nature. A face image $\boldsymbol{y}$ can be expressed as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{k}$, where $\boldsymbol{k}$ represents the sparse coefficient vector, and $\boldsymbol{X}$ is a data matrix whose columns represent the training data. It should be known that the samples of the same subject are highly correlated, while the correlation becomes weak when samples are from different subjects. For the samples in $\boldsymbol{X}$ belonging to the same subject of $\boldsymbol{y}$, the corresponding coefficients in the sparse coefficient vector $\boldsymbol{k}$ should have non-zero values, while the rest of the coefficients are zero. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_M]$, where $\boldsymbol{x}_j$ is the j^{th} training sample and M is the total number of training samples. Let us consider that there are c classes in the training set, and n samples for each class, i.e. $M = nc$. According to sparse theory, each training sample can be linearly reconstructed by the remaining $M - 1$ samples, with most of the weights of the samples being zero. Our major objective is to project the features of training and testing samples into a low-dimensional feature space, such that their sparsity is preserved. Let us assume that the sparse coefficient vectors of the training samples are denoted as $\boldsymbol{K} = [\boldsymbol{k}_1, \boldsymbol{k}_2, \dots, \boldsymbol{k}_M]$, where $\boldsymbol{k}_j \in R^M$ is the sparse vector of the j^{th} training sample, computed using the l_1-minimization technique.

3.5.1 Feature Representation based on Sparse Coding

The purpose of sparse representation is to represent test images by using the minimum number of training samples. Mathematically, it can be written as follows:

$$\min\|\boldsymbol{k}\|_0 \text{ s.t. } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{k}. \tag{3.17}$$

If it contains enough sparsity, then the solution of equation (3.17) is same as solving the l_1-minimization problem, i.e.

$$\min\|\boldsymbol{k}\|_1 \text{ s.t. } \boldsymbol{y} = \boldsymbol{X}\boldsymbol{k}. \tag{3.18}$$

In an ideal situation, a test image $\boldsymbol{y}$ from the j^{th} class can be linearly represented in terms of all the training samples, which can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{k} = \boldsymbol{X}\boldsymbol{k}_{1j} + \boldsymbol{X}\boldsymbol{k}_{2j} +, \dots, +\boldsymbol{X}\boldsymbol{k}_{nj}, \tag{3.19}$$

where n is the number of training samples in the j^{th} class, and $\boldsymbol{k}_{ij}$ is the sparse coefficient vector whose entries are non-zero for the ones associated with the j^{th} class and the i^{th} training sample in the class.

In our algorithm, we propose the following objective function to compute the optimal projection matrix $\boldsymbol{P}$, which preserves the sparse structure when projecting the extracted multiple low-rank features $\boldsymbol{F}$ onto a new feature subspace.

$$\boldsymbol{P} = \arg\min_{\boldsymbol{P}} \sum_{j=1}^{M} \left\| \boldsymbol{P}^T \boldsymbol{f}_j - \boldsymbol{P}^T \boldsymbol{F}\, \boldsymbol{k}_j \right\|^2, \tag{3.20}$$

where $\boldsymbol{f}_j$ is the low-rank feature vector of the j^{th} training sample. By using simple algebraic formulation, Equation (3.20) can be written as:

$$\min \boldsymbol{P}^T \left(\sum_{j=1}^{M} (\boldsymbol{f}_j - \boldsymbol{F}\boldsymbol{k}_j)(\boldsymbol{f}_j - \boldsymbol{F}\boldsymbol{k}_j)^T \right) \boldsymbol{P}. \tag{3.21}$$

Assume that the low-rank feature vectors are projected onto an m dimensional vector space. Let $\boldsymbol{u}_j$ be the m-dimensional unit vector with the j^{th} element equal to 1, and 0 otherwise. Equation (3.21) can then be written as follows:

$$\begin{aligned}
&\min \boldsymbol{P}^T \left(\sum_{j=1}^{M} (\boldsymbol{F}\boldsymbol{u}_j - \boldsymbol{F}\boldsymbol{k}_j)(\boldsymbol{F}\boldsymbol{u}_j - \boldsymbol{F}\boldsymbol{k}_j)^T \right) \boldsymbol{P} \\
&= \min \boldsymbol{P}^T \boldsymbol{F} \left(\sum_{j=1}^{M} (\boldsymbol{u}_j - \boldsymbol{k}_j)(\boldsymbol{u}_j - \boldsymbol{k}_j)^T \right) \boldsymbol{F}^T \boldsymbol{P} \\
&= \min \boldsymbol{P}^T \boldsymbol{F} \left(\sum_{j=1}^{M} (\boldsymbol{u}_j \boldsymbol{u}_j{}^T - \boldsymbol{u}_j \boldsymbol{k}_j{}^T - \boldsymbol{k}_j \boldsymbol{u}_j{}^T + \boldsymbol{k}_j \boldsymbol{k}_j{}^T) \right) \boldsymbol{F}^T \boldsymbol{P} \\
&= \min \boldsymbol{P}^T \boldsymbol{F} (\boldsymbol{I} - \boldsymbol{K} - \boldsymbol{K}^T + \boldsymbol{K}^T \boldsymbol{K}) \boldsymbol{F}^T \boldsymbol{P}.
\end{aligned} \tag{3.22}$$

We set the constraint $\boldsymbol{P}^T \boldsymbol{F} \boldsymbol{F}^T \boldsymbol{P} = 1$. Then, the objective function is converted into the following optimization problem

$$\min_{\boldsymbol{P}} \frac{\boldsymbol{P}^T \boldsymbol{F} (\boldsymbol{I} - \boldsymbol{K} - \boldsymbol{K}^T + \boldsymbol{K}^T \boldsymbol{K}) \boldsymbol{F}^T \boldsymbol{P}}{\boldsymbol{P}^T \boldsymbol{F} \boldsymbol{F}^T \boldsymbol{P}}. \tag{3.23}$$

To solve (3.23), the Lagrange method is used, which provides the following equation:

$$L(\boldsymbol{P}, \lambda) = \boldsymbol{P}^T \boldsymbol{F} (\boldsymbol{I} - \boldsymbol{K} - \boldsymbol{K}^T + \boldsymbol{K}^T \boldsymbol{K}) \boldsymbol{F}^T \boldsymbol{P} - \lambda (\boldsymbol{P}^T \boldsymbol{F} \boldsymbol{F}^T \boldsymbol{P} - 1), \tag{3.24}$$

where λ is a lagrange multiplier, and $\boldsymbol{I}$ represents the identity matrix. To compute the optimum sparse projection matrix $\boldsymbol{P}$, we set the derivative to zero, i.e. $\frac{\partial L}{\partial \boldsymbol{P}} = 0$, which gives the following equation:

$$\boldsymbol{F}(\boldsymbol{I} - \boldsymbol{K} - \boldsymbol{K}^T + \boldsymbol{K}^T\boldsymbol{K})\boldsymbol{F}^T\boldsymbol{P} = \lambda \boldsymbol{F}\boldsymbol{F}^T\boldsymbol{P}. \tag{3.25}$$

Finally, it becomes an eigen-decomposition problem in which we select the m eigenvectors of the matrix $(\boldsymbol{F}\boldsymbol{F}^T)^{-1}\boldsymbol{F}(\boldsymbol{I} - \boldsymbol{K} - \boldsymbol{K}^T + \boldsymbol{K}^T\boldsymbol{K})\boldsymbol{F}^T$, with the smallest eigenvalues to construct a new low-dimensional feature subspace. Our proposed algorithm builds the sparse coefficient matrix by utilizing all the training data, so no search for nearest neighbors is required during testing. For visualization of the learned low-rank sparse features, we randomly selected 10 face samples from each of the 10 different classes. The low-rank sparse features are first extracted and then visualized using t-Distributed stochastic neighbor embedding (t-SNE) [168], as shown in Fig. 3-5. It can be observed that the discriminability of the learned features is enhanced, as samples from different classes are well separated in the feature space.

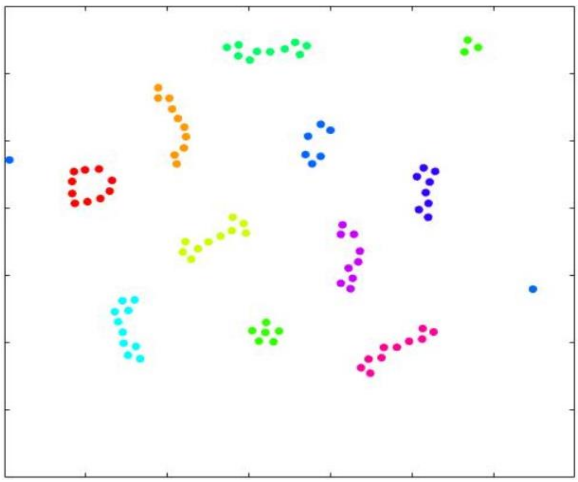


Fig. 3-5. Visualization of the learned low-rank sparse features using t-SNE.

3.5.2 Geometrical and Mathematical properties of a Generalized Eigenvalue Problem

In this section, we explain our proposed formulation from geometrical and mathematical points of view. The proposed formulation involves two major steps, which are: (1) construction of the sparse coefficient matrix $\boldsymbol{K}$, and (2) determination of the projection matrix $\boldsymbol{P}$. To understand this eigenvalue problem, attention needs to be paid to these two major steps. Firstly, sparse coefficient vector $\boldsymbol{k}_j$ is determined for each sample $\boldsymbol{x}_j$ using l_1 minimization. Now, we will analyze the effectiveness of this computed sparse coefficient matrix $\boldsymbol{K}$. Geometrically, each sparse coefficient vector $\boldsymbol{k}_j$ is invariant to scaling and rotation of the data samples. It is also invariant to translation due to the constraint

$1 = \mathbf{1}^T \boldsymbol{k}_j$, where $\mathbf{1}^T$ is an identity vector. Therefore, the sparse coefficient matrix $\boldsymbol{K}$ remains unchanged whenever data samples are translated and rotated, which is one of its important geometrical properties. In our proposed method, we construct a coefficient matrix, using whole training data, instead of using the k-nearest neighbors. This helps in preserving the global structure of the data, while projecting them into a new sparse feature subspace. The constructed sparse coefficient matrix also has the capability to preserve the discriminant information. To understand this, let us take an example related to face recognition. It is assumed that the face images, belonging to the same class, lie on a linear subspace. Let $\boldsymbol{x}_j$ be a face image belonging to the j^{th} class, $\boldsymbol{x}_j$ can be represented as a linear combination of the other face images from the same j^{th} class, and the computed coefficient vector $\boldsymbol{k}_j$ is sparse. This shows that $\boldsymbol{k}_j$ naturally contains discriminant information, so it can easily distinguish the face images of two different classes. Fig. 3-5 shows the training stage of our proposed framework.

3.5.3 Useful Properties

Our proposed sparse-coding-based method exhibits some major useful properties, which are as follows:

- Our proposed method uses l_1 regularization, which enables it to encode the prior knowledge of sparsity, resulting in the extraction of more discriminative information from the data. It performs sparse reconstruction only in the training process. Having determined the projection matrix $\boldsymbol{P}$, sparse reconstruction is no longer necessary and our algorithm is therefore efficient.
- Samples from two different classes may have significant overlap in the subspaces obtained by using PCA, NPE, and LPP. Usually, PCA suffers the most, because the eigenvectors selected are those that best represent the data, rather than distinguishing them. To solve this issue, LPP and NPE attempt to consider the local properties (structure) of the data. However, the real structure cannot be identified by using a pre-defined neighbourhood size. In comparison to them, our proposed method can separate face images from different classes in an accurate way. This can be better explained using the idea of sparsity, which assumes that each data point can be better represented by as small number of samples as possible. The learned projection matrix has a high-degree of sparseness, greater than 80%. We found that high degree of sparsity is quite useful in feature learning, but do not have a severe impact on the classification performance. As the sparsity decreases, the recognition rate drops gradually by 3-5%, and then remains unchanged.

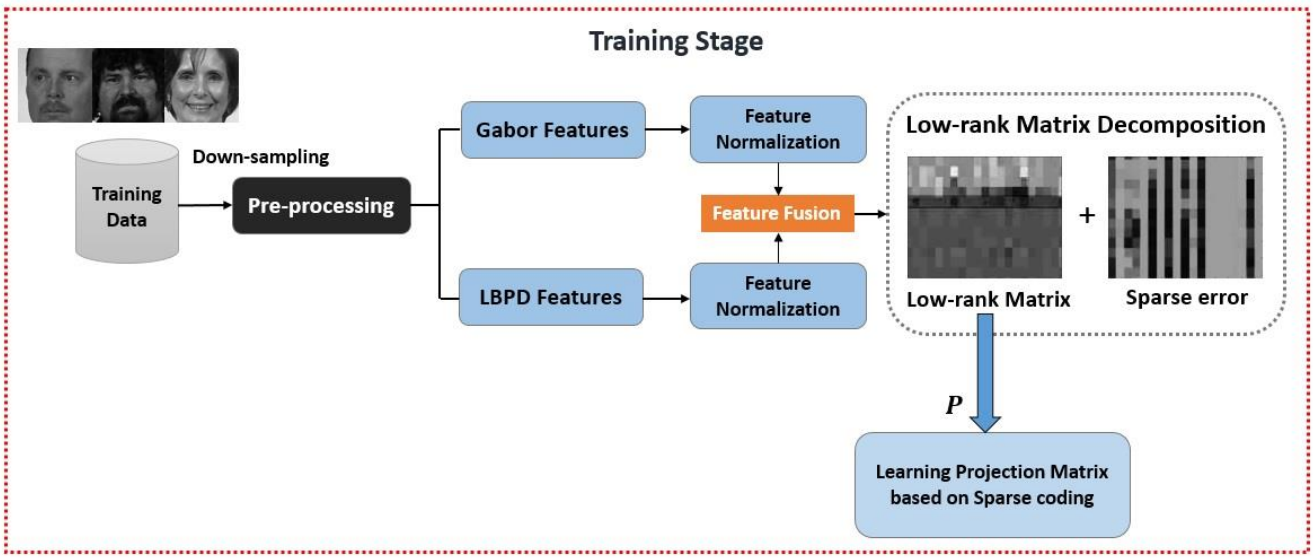


Fig. 3-6. Training stage of our proposed framework.

3.6 Linear-Regression-based Classification

After projecting gallery and probe features onto the sparse feature subspace, a linear regression framework [43] is utilized to model the similarity between them. It computes a linear mapping function between the gallery and the probe face images. According to the developed model, a linear relationship exists between a probe image and all the images in a gallery set. If a query face image fits to the i^{th} class in the gallery set, it can be expressed linearly in terms of the gallery-images features from the same class. Therefore, we have

$$\boldsymbol{I}_R = \boldsymbol{X}_i \boldsymbol{\alpha}_i \,, \tag{3.26}$$

where $\boldsymbol{I}_R$ is the reconstructed probe image based on the gallery images from the i^{th} class; $\boldsymbol{X}_i = [\boldsymbol{x}_{i,1}, \dots, \boldsymbol{x}_{i,n_i}]$ are the training samples from the i^{th} class, which has n_i samples; and $\boldsymbol{\alpha}_i$ represents the image coefficient vector of the probe image, estimated by the least-squares algorithm. The next step is to find the residual values for each class, based on the computed coefficient vectors. The query face image $\boldsymbol{y}$ is assigned to the class j which has the minimum residual value, i.e.

$$j = \min_i \|\boldsymbol{y} - \boldsymbol{X}_i \boldsymbol{\alpha}_i\|. \tag{3.27}$$

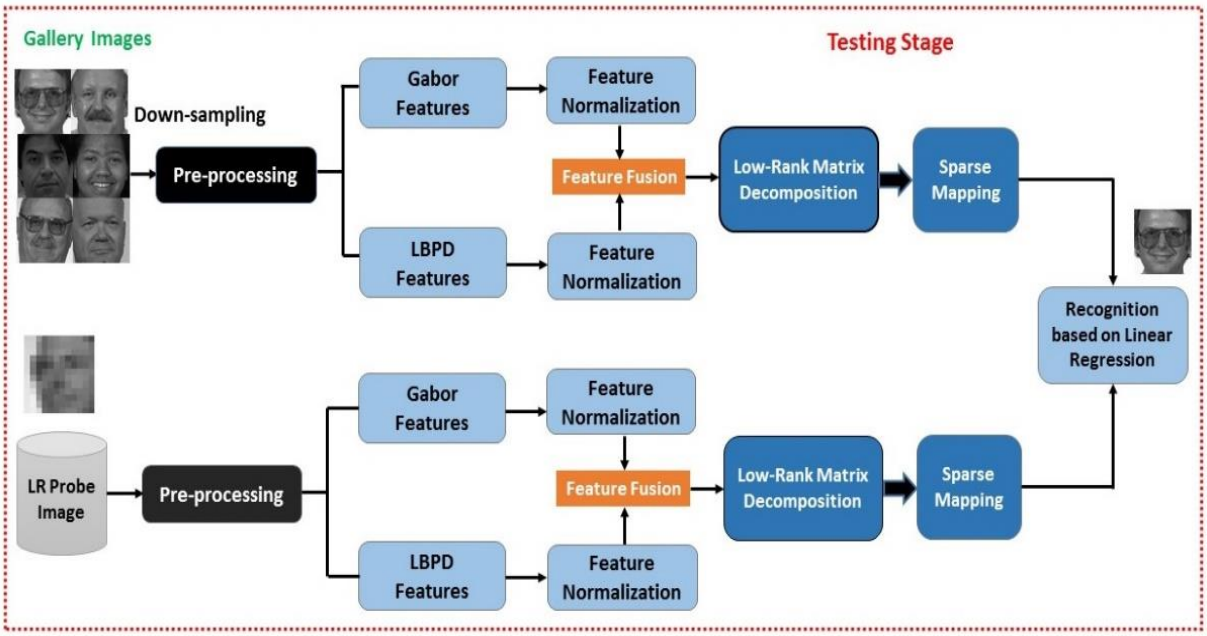


Fig. 3-7. Testing stage of the proposed framework.

3.7 Experiments

To evaluate the effectiveness of our proposed approach, we conduct extensive sets of experiments on five face datasets, which include Extended Yale-B, Multi-PIE [169], FERET [170], LFW [15], and Remote Face [171] databases. In the pre-processing stage, face images are first detected and aligned using MTCNN [21]. For the different databases, we followed the other papers on LR face recognition to down-sample face images to a specific size. This allows our algorithm to be directly compared to other algorithms.

3.7.1 Experimental Results on the Extended Yale-B Dataset

The Extended Yale-B dataset consists of 2,432 images from 38 subjects with 64 images per subject, which were taken with different illumination conditions. In our experiments, all 64 images per subject with different illumination conditions are utilized. For training, we randomly select 10, 20, and 30 images per subject. LR probe images of size 12×12 are generated using a down-sampling operation. The HR and LR face images of five individuals from the dataset are shown in Fig. 3-9 (a). In Fig. 3-8, the recognition rate shows an increasing trend and remains stable as feature dimension increases. Our proposed method performs better than the other LR face recognition methods, and achieves the highest recognition rate of 94.74%, when the feature dimension is higher than 70. The highest recognition rates of CLPM [10], DSR [9], NMCF [69], MDS [73], CCA [77], CDMMA [78], and CMDA [80] are 89.35%, 62.3%, 81.42%, 77.01%, 77.90%, 88.2%, and 89.8%, respectively, at their corresponding optimal feature dimensions.

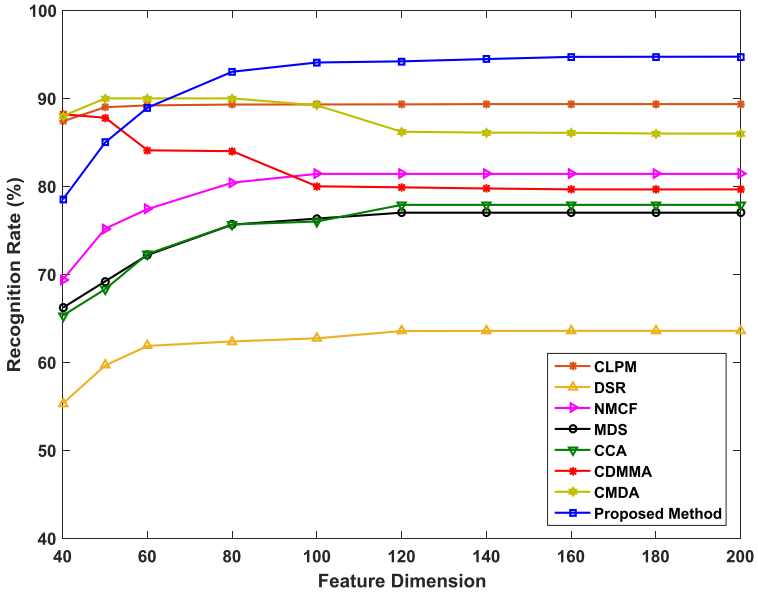


Fig. 3-8. Recognition rates of different methods at different feature dimensions on the Extended Yale-B database (LR: 12×12).

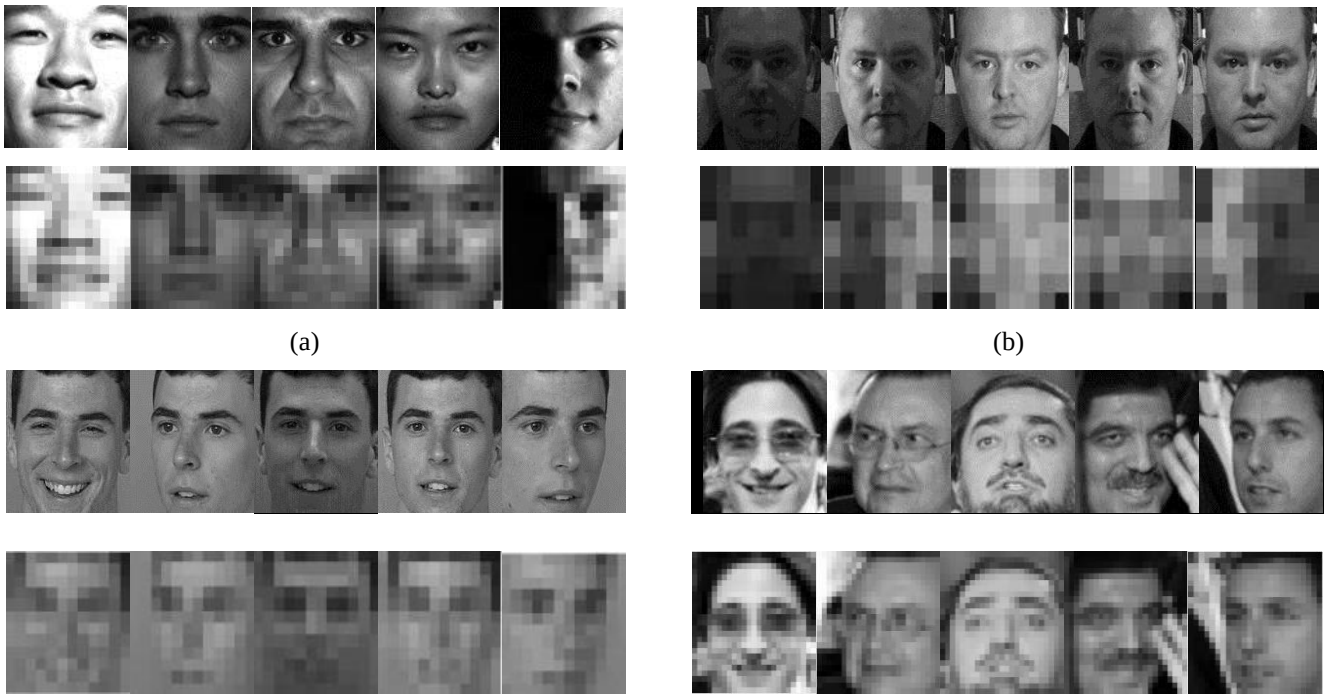


Fig. 3-9. Original images and the corresponding LR images: (a) Extended Yale-B, (b) Multi-PIE, (c) FERET, and (d) LFW databases. The first rows show the original face images, while the second rows show the downsampled images.

3.7.2 Experimental Results on the Multi-PIE Dataset

The Multi-PIE dataset consists of more than 700,000 face images of 337 persons. Images were captured in four different sessions. Following the protocol used in [75], we conducted experiments on a subset of session 04, which

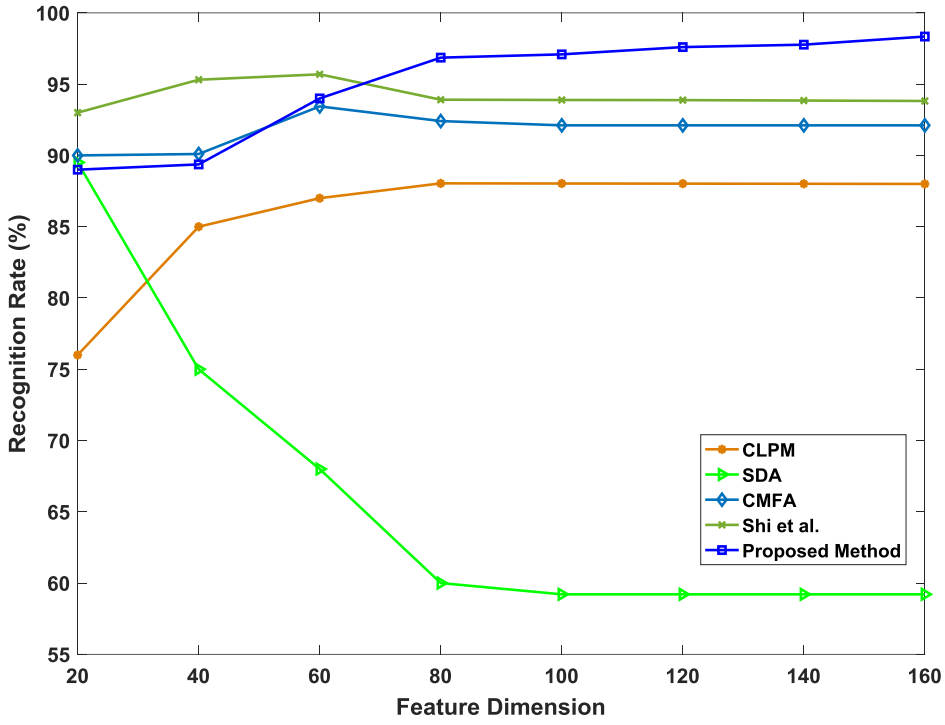
contain images with frontal pose under 20 different illumination conditions. The camera and the recording numbers, used in our experiments, are 05-1 and 01, respectively. For training, we randomly select 50 subjects, while the rest of the subjects are used for testing. To construct a gallery set, 6 images of each subject are selected, while the remaining 14 images are included in a probe set. LR probe images of size 8×8 are generated using a down-sampling operation. The sample HR and the LR images from the Multi-PIE dataset are shown in Fig. 3-9 (b). Table 3-2 shows the comparative results in terms of Rank-1 recognition accuracy. Fig. 3-10(a) shows the recognition rate at different feature dimensions. Our method outperforms the other LR face recognition methods and achieves the highest recognition accuracy of 97.41%.

Table 3-1. Comparative results on the FERET (Fa) dataset, in terms of Rank-1 Recognition accuracy, at different resolutions with optimal feature dimensions.

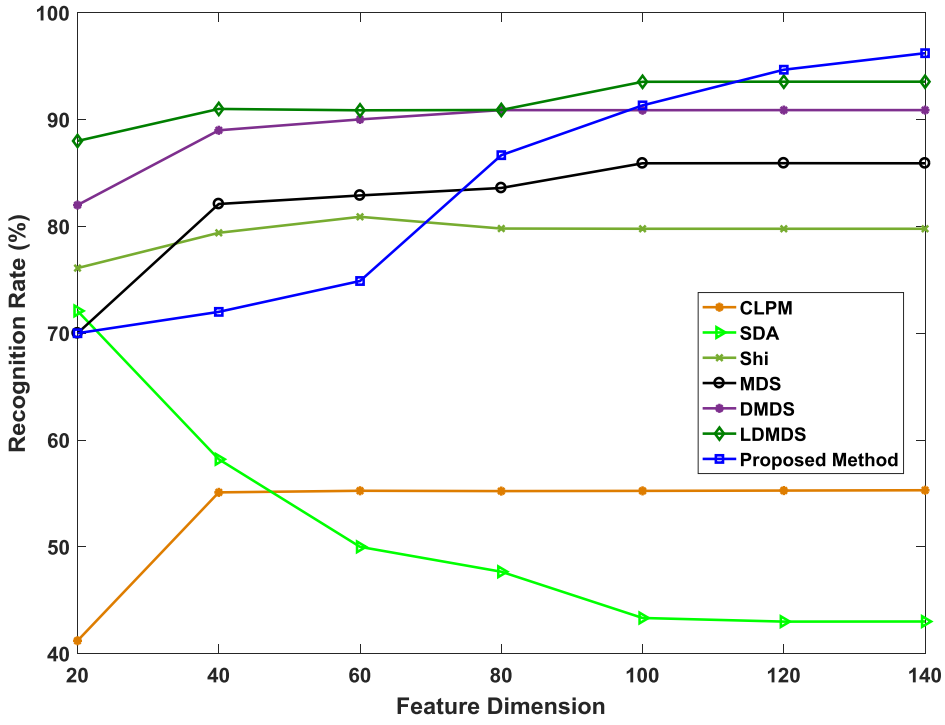
Algorithm	8*8	12*12	16*16
CLPM [10]	79.94%	82.46%	84.48%
CMFA [74]	72.08%	75.40%	75.60%
SDA [72]	68.75%	71.77%	72.08%
C-RSDA [79]	82.36%	86.29%	86.29%
Proposed method	**84.71%**	**89.66%**	**95.22%**

Table 3-2. Comparative results for FERET (BaBe) and Multi-PIE datasets in terms of Rank-1 Recognition accuracy at optimal feature dimensions (Probe image resolution:8×8).

Algorithm	FERET (BaBe)	Multi-PIE
CLPM [10]	55.22%	88.04%
CMFA [74]	75.98%	93.44%
SDA [72]	72.09%	89.51%
Shi et al. [75]	80.90%	95.69%
MDS [73]	85.91%	91.78%
LMCM [76]	90.00%	--------
DMDS [81]	90.89%	93.88%
LDMDS [81]	93.55%	95.81%
Proposed Method	**96.22%**	**97.41%**



(a)



(b)

Fig. 3-10. Recognition rates with different feature dimensions: (a) Multi-PIE database (LR: 8 × 8) and (b) the FERET database (BaBe) (LR: 8 × 8).

3.7.3 Experimental Results on the FERET Dataset

The FERET dataset is one of the widely used face datasets for performance evaluation. It consists of more than 13,000 face images from 1,565 subjects. The images of each subject are taken with variations in illumination, expressions, and pose. Performance on FERET dataset is evaluated using three probe sets (Fb, Dup1, and Dup2), against one standard gallery set (Fa), respectively. The HR and LR face images of five individuals from the FERET dataset are shown in Fig. 3-9(c). In our experiments, we select a group, namely Fa, which consists of 994 frontal face images with one image per subject, and used it as a gallery set, while Fb consists of 994 images with expression variations as a probe set. Throughout this chapter, we called this subset 'FERET (Fa)'. We carried out our experiments by selecting a training set that consists of only one image per subject. LR probe images of size, 12×12, are generated by using a down-sampling operation. Our method outperforms other LR face recognition methods, and achieves the recognition rate of 89.66%, with a feature dimension of 200, whereas the recognition accuracy achieved by CLPM [10], CMFA [74], SDA [72] and C-RSDA [79] are 82.46%, 75.40%, 71.77%, and 86.29%, respectively, with their optimal feature dimensions. Table 3-1 shows the comparative results at different resolutions.

For comprehensive analysis, we further perform experiments on another challenging subset of the FERET dataset, which contains images of 200 subjects having large variations in pose, and expression (including $ba, bd, be, bf, bg, bj\ and\ bk$). Throughout this chapter, we called this subset 'FERET (BaBe)'. In this subset, 7 images per subject are available. In our experiments, 50 subjects are selected for training, while the remaining 150 subjects are used to construct a testing set. During training, all the 7 images per subject are used. For the testing set, the first four images are used to construct a galley set, while the remaining 3 images are used as a probe set. For evaluation, we down-sample the probe images to the size of 8×8. Our method achieves the recognition rate of 96.22% with a feature dimension of 140, which is better than the other LR face recognition methods. The comparative results are shown in Table 3-2. Recognition rate with different feature dimensions are shown in Fig. 3-10 (b).

3.7.4 Experimental Results on the LFW Database

Recognizing face images taken under unconstrained environments is more challenging. To do so, we conduct experiments on the LFW dataset. All the images in this dataset were captured in the wild, having large variations in expression, pose, make-up, lighting condition, etc. It consists of 13,233 face images from 5,749 individuals. Out of these, 1,680 individuals contain more than two images, and 610 of them contain more than four images in the dataset.

We randomly select 4 images from each of the 610 individuals. For training, we randomly select 150 subjects with 10 images each from the CASIA-Web face dataset [57]. For testing, two images per subject from selected LFW images are used to construct the gallery set, and the other two are used for the probe set. LR probe images of resolutions 12×12, 16×16, and 20×20 are generated using a down-sampling operation. Our method shows a promising result by achieving 88.23% accuracy on LR images of size 12×12. For the LFW dataset, we compare our results with a recently proposed deep-learning-based method [49], which will be shown in the next section. The HR and LR face images of five individuals from the LFW database are shown in Fig. 3-9(d). The recognition rate of our proposed method, with different feature dimensions and at three different probe image resolutions, is shown in Fig. 3-11.

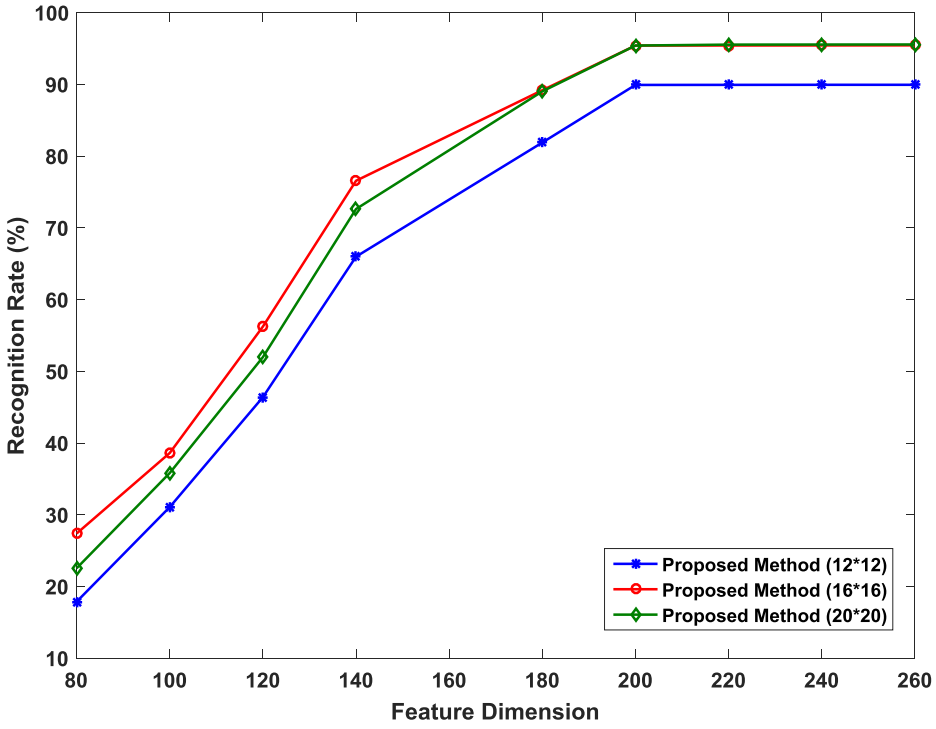


Fig. 3-11. Recognition rates of our proposed method on the LFW database, with different feature dimensions and at different probe image resolutions.

3.7.5 Comparison with Deep-Learning Methods

Convolutional neural networks (CNNs) have revolutionized pattern-recognition research by providing superior performances in various machine learning tasks. One of the main reasons for its success is the availability of a large amount of training data and the networks are trained for feature extraction and recognition from end to end. As discussed earlier, deep-learning methods have achieved more than 99% recognition accuracy on the LFW dataset. However, deep learning is still finding a way to make its mark in solving the LR face recognition problem. Schroff et

al. [6] reported around 50% decline in validation rate when the size of face images is reduced from 256×256 to 40×40. Similarly, Chevalier at al. [172] also reported the significant decline in recognition performance of the CNN models, by varying the image resolution from 100×100 to 20×20. Now, we will analyze the performance of a deep-learning-based approach for recognition of LR face images. Then, we analyze the performance of our proposed approach in comparison to the deep-learning-based method. In this regard, we have conducted several experiments to evaluate whether current deep face models are good enough for recognizing LR face images.

For deep-learning-based experiments, we used three different models of SphereFace [49] (a deep CNN model, trained on the CASIA-Web Face [57] dataset, to perform face recognition). It is worth noting that the size of the training data used by SphereFace is small as compared to the datasets used in VGG-Face [48], FaceNet [6], and Deep Face [3]. The model achieves excellent performance on both LFW [15] and YouTube face (YTF) [173] datasets. First, we used a pretrained model (Model no. 1) of Sphere Face to perform LR face recognition. Our experimental results show that using a pretrained model for LR face recognition gives the worst performance. The reason for this is that SphereFace is originally trained on HR images, so fine-tuning or retraining is necessary to achieve optimal performance on LR face images. It should be noted that the LFW and Multi-PIE datasets contain color face images. Therefore, we fine-tune the model with a small learning rate of 0.01, which linearly decay to zero. We also randomly down-sample the input face images to the sizes of $8 \times 8 - 95 \times 95$, and report the best performance for both datasets in Tables 3-4 and 3-7, respectively. On the other hand, the Extended Yale-B and FERET datasets consist of grayscale images, so fine-tuning the original model with color images might not give the optimal performance on these two datasets. Therefore, we retrain the original model using gray-scale images instead of color face images. In this case, the input size of the first convolutional layer is changed to 1, rather than 3. Similar to fine-tuned models, the input face images are down-sampled to the size of $8 \times 8 - 95 \times 95$. Initially, we set the learning rate to 0.1, which linearly decays to zero. The optimal recognition performances are reported in Tables 3-3, 3-5, and 3-6, respectively.

Table 3-3. Recognition results of deep-learning-based methods and the proposed method for LR face recognition at different resolutions, on the Extended Yale-B database.

Method	12×12	16×16	20×20
SphereFace-HR	61.22%	54.64%	57.12%
SphereFace-LR	63.08%	68.03%	73.61%
Proposed Method	**94.74%**	**95.13%**	**94.87%**

Table 3-4. Recognition results of deep-learning-based methods and the proposed method for LR face recognition at different resolutions, on the Multi-PIE Dataset.

Method	8 × 8	12 × 12	16 × 16	20 × 20
SphereFace-HR	84.39%	90.59%	96.03%	99.02%
SphereFace-LR	96.94%	98.41%	99.96%	**100.00%**
Proposed Method	**97.41%**	97.94%	97.94%	98.06%

Table 3-5. Recognition results of deep-learning-based methods and the proposed method for LR face recognition at different resolutions, on the FERET (Fa) database.

Method	8 × 8	12 × 12	16 × 16	20 × 20
SphereFace-HR	52.73%	65.07%	72.30%	77.64%
SphereFace-LR	59.29%	49.28%	86.76%	94.22%
Proposed Method	**84.71%**	**89.66%**	**95.22%**	**96.55%**

Table 3-6. Recognition results of deep-learning-based methods and the proposed method for LR face recognition at different resolutions, on the FERET (BaBe) database.

Method	8 × 8	12 × 12	16 × 16	20 × 20
SphereFace-HR	42.67%	65.56%	80.22%	88.44%
SphereFace-LR	55.11%	54.67%	94.22%	99.56%
Proposed Method	**96.22%**	**99.33%**	**98.67%**	**99.78%**

Table 3-7. Recognition results of deep-learning-based methods and the proposed method for LR face recognition at different resolutions, on the LFW database.

Method	12 × 12	16 × 16	20 × 20
Sphere-Face-HR	18.93%	19.67%	41.64%
Sphere-Face-LR	52.70%	78.85%	89.92%
Proposed Method	**88.36%**	**93.28%**	**95.41%**

It can be seen that fine-tuning and retraining produce significantly major improvements in the recognition of LR face images. Throughout our experiments, we follow the Sphere Face's implementation to align the face images using MTCNN [21]. In the testing stage, we extract deep features from the FC1 layer. In the experiments, the final feature of a probe face image is obtained by concatenating the original features and its horizontally flipped version. Finally, the similarity score is computed using the cosine similarity. In [49], the performance of the SphereFace model was evaluated using different numbers of layers (10, 20, 36, 40 and 64). However, a minor improvement is reported in the

recognition rate when the number of layers is increased from 20 to 64. Throughout our experiments, we first down-sample the gallery faces to the same resolution as the probe image to achieve better performances.

(a)

(b)

Fig. 3-12. Matching results of our proposed approach, with the probe images shown on the left: (a) matching under large pose variation, and (b) matching under expression and lighting variations.

3.7.6. Experimental Results on the Remote Face Database

In order to evaluate the performance of our proposed method under more challenging conditions, we conducted experiments on the Remote Face dataset [171], which contains images taken under unconstrained outdoor maritime environments. The images were taken at different distances, ranging from 10-250 m. There are 2,102 face images in total from 17 subjects. Each subject has a number of images, ranging between 29 and 306.

The dataset consists of six subsets, denoted as *blur*, *illum*, *illum_blur*, *frontal_pose*, *Nf_pose*, and *low_res*, respectively. The *blur* subset contains 75 face images with blurring effects. The *illum* subset consists of 561 face images with different lighting conditions. The *illum_blur* subset consists of 128 images, with both lighting and blurring effects. The *low_res* subset is the most challenging one, as it contains 90 face images of very low resolution. The *frontal_pose* and non-frontal pose (*Nf_pose*) subsets include images having frontal and non-frontal poses, with 1,166 and 846 face images, respectively. The gallery set consists of five HR face images of each subject. We conducted experiments on all of the six subsets and achieved competitive results. For the subsets with a blur and lighting effects, all the methods can achieve promising results, as these effects do not have severe impact on the image's appearance. All the methods

can also achieve satisfactory performance on the frontal-pose subset. For the *Nf_pose* subset, most of the methods have their performance drop significantly, while our method can achieve the best performance, with the recognition rate of 96.4%. The *low_res* subset is the most challenging one, because the images are of a small size and suffered from blurring. The resolution of the images in this subset is 20×30 only. The face images of three different subjects from all the six subsets are shown in Fig. 3-13.

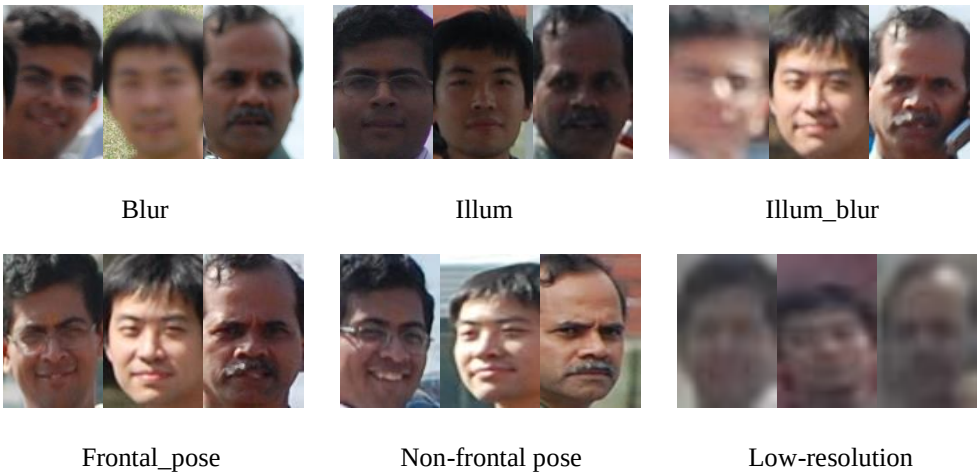


Fig. 3-13. Sample face images from all the six subsets of the Remote Face database.

No training set is provided by the Remote Face dataset [171], so we used 16,028 frontal face images from the Face Recognition Grand Challenge (FRGC) dataset for training. It can be seen that all the other methods do not perform well on the LR face images, as it is very difficult to extract useful information from such low-quality, low-resolution images. Similar to [36], we compare the performance of our method with different local feature descriptors, including binarized statistical image features [174], DFD [34], LSF [175], and two deep-learning-based methods (DL [176], and SphereFace (SF) [32]). Experimental results are tabulated in Table 3-8.

It can be seen that the deep-learning-based model SF [32] can achieve state-of-the-art performance on the five subsets, except the *low_res* subset. This is because the original model was trained on HR face images. In our experiments, we have also fine-tuned the model using LR images from the CASIA-Web Face dataset [66], with the sizes between 8×8 and 95×95, for performance evaluation. In the experiment results, the fine-tuned model is denoted as SF-FT. It is worth noting that the original SF model performs better than the fine-tuned model (SF-FT), because the distribution of the downsampled faces used for fine-tuning is different from the native LR faces. It can be observed that the performance of the deep learning models declines when the presented probe images are of low resolution. However, our method performs better than all the other methods on this subset, as well as on the

frontal_pose, *illum_blur*, and *Nf_pose*. For the *blur* and *illum* subsets, the performance of our method is comparable to SF and SF-FT.

Table 3-8. Comparative results on Remote Face dataset, in terms of Rank-1 Recognition rate on all the six subsets.

Subset	Algorithm	Rate (%)	Subset	Algorithm	Rate (%)	Subset	Algorithm	Rate (%)
blur	BSIF	62.2	*frontal_pose*	BSIF	70.1	*Nf_pose*	BSIF	49.2
	DFD	63.5		DFD	78.6		DFD	52.5
	DL	48.6		DL	80.3		DL	49.8
	Shearlet	62.5		Shearlet	77.8		Shearlet	51.7
	LSF	67.3		LSF	83.8		LSF	57.2
	SF-FT	89.4		SF-FT	95.4		SF-FT	90.1
	SF	**93.8**		SF	97.8		SF	94.2
	Ours	90.5		Ours	**98.8**		Ours	**96.4**
illum	BSIF	79.3	*illum_blur*	BSIF	74.8	*low_res*	BSIF	11.2
	DFD	83.4		DFD	75.2		DFD	14.5
	DL	80.4		DL	71.8		DL	11.5
	Shearlet	81.6		Shearlet	74.3		Shearlet	13.8
	LSF	92.5		LSF	76.0		LSF	19.9
	SF-FT	**99.2**		SF-FT	96.9		SF-FT	51.4
	SF	**99.4**		SF	98.5		SF	66.0
	Ours	98.8		Ours	**98.8**		Ours	**81.1**

3.7.6 Feature Fusion

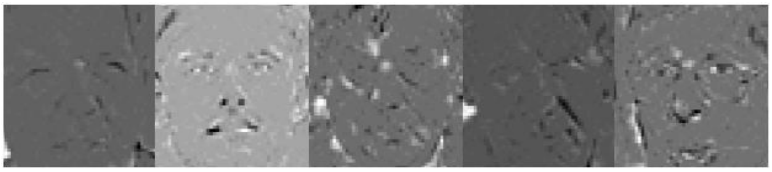
Most of the existing face recognition methods utilize only one feature descriptor. However, in difficult tasks such as LR face recognition, no single feature is good enough to extract all the relevant information from LR face images. Combining multiple efficient features is a promising way to bring major improvement in recognition accuracy. In our previous approach [177], we used only one local feature descriptor (Gabor wavelets) for extracting facial details from LR images, which provided satisfactory performance. However, it does not perform well under large pose variations and poor lighting conditions.

Table 3-9. Recognition rates in comparison to the preliminary work [177], recorded at the optimal feature dimensions.

Dataset	Previous Method [177]	Proposed Method
Extended Yale-B (LR: 12 × 12)	49.21%	**94.74%**
FERET (fa) (LR: 12 × 12)	70.08%	**89.66%**
FERET (BaBe) (LR: 8 × 8)	46.44%	**96.22%**
LFW (LR: 12 × 12)	44.34%	**88.36%**
Multi-PIE (LR: 8 × 8)	56.40%	**97.41%**

To overcome this problem, we fuse two efficient local feature descriptors, i.e. Gabor wavelets and LBPD, which can achieve much better performance even under large unconstrained environments. Table 3-9 shows the recognition rates on the four databases (in comparison to previously proposed approach [177]), with and without performing feature fusion at the corresponding optimal feature dimensions. It can be observed that learning and fusing the low-rank features brings significant improvements, in terms of recognition accuracy. Two matching results are shown in Fig. 3-12, with pose, expression and lighting variations. Recognition results based on our method recorded for all the four datasets, at the corresponding optimal feature dimensions, using different numbers of training samples per subject are shown in Tables 3-10, 3-11, 3-12, and 3-13, respectively. It should be noted that coupled mapping methods operate directly on data matrices by projecting HR training and LR testing samples onto a common feature subspace, without extracting discriminant information from the facial images. Our proposed method not only utilizes features robust to resolution, but also employs sparse coding, which preserves the sparsity of the data samples. This also makes our method suitable for recognition of very LR images, even down to the size of 8×8. As discussed before, we decompose the extracted local features into a low-rank feature matrix, and a sparse error matrix. After that, only low-rank component is utilized for identification. In this section, we evaluate the performance of our proposed method with and without including the estimated sparse error matrix in the final feature representation. Firstly, we visualize the estimated low-rank and sparse error components by applying the low-rank matrix decomposition algorithm on some face images from the LFW dataset. Fig. 3-14 shows the low-rank representation of face images and their corresponding sparse errors.

The low-rank representation of face images from LFW dataset.



Corresponding Sparse error images

Fig. 3-14. Low-rank representation of face images and the corresponding sparse error images.

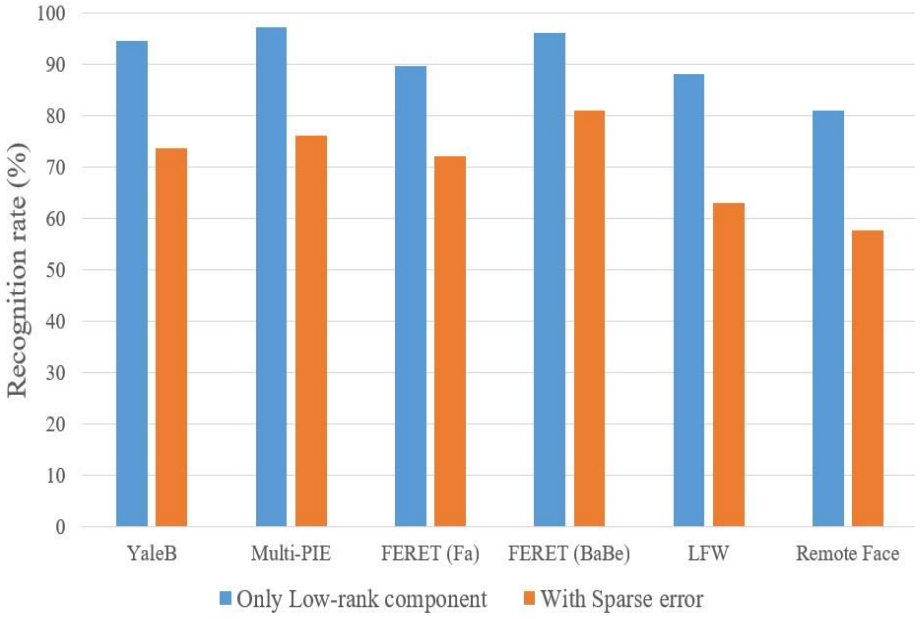


Fig. 3-15. Recognition rates with and without using the sparse error matrix for all the five datasets at optimal feature dimensions.

In our experiments, we found that by including the sparse error matrix in the final feature representation, recognition rate drops by a significant level. We repeat our experiments by following the protocol discussed before, and report the recognition rates for all of the five datasets at optimal feature dimensions. Fig. 3-15 shows the recognition rates with and without using the sparse error matrix. It can be observed that the recognition rate increases by 20-25%, when the sparse error matrix is discarded.

3.7.7 Recognition across Different Probe Resolutions

In this section, the performance of our proposed method is evaluated using probe images of different resolutions, with and without using our proposed morphological preprocessing method. For all the four datasets, three different probe resolutions, 8×8, 12×12, and 16×16, were used. Experiments were conducted by selecting a fixed number of training images per subject and the respective optimal feature dimensions, at different probe resolutions, and results are reported in Fig. 3-16.

The results prove that our proposed method can achieve a very good recognition performance even if image resolution is reduced to lower than 12×12. However deep-learning-based methods are not optimistic for recognizing the LR images of size 8×8. As discussed earlier, pre-processing is considered as one of the most important steps prior to face-recognition, so we also compute the recognition rates with and without using the proposed morphological pre-processing method, at the corresponding optimal feature dimensions for the four datasets. The corresponding results

are shown in Table 3-14. It can be observed that 1% to 6% of improvement, in terms of recognition rate, can be obtained when the pre-processing step is employed.

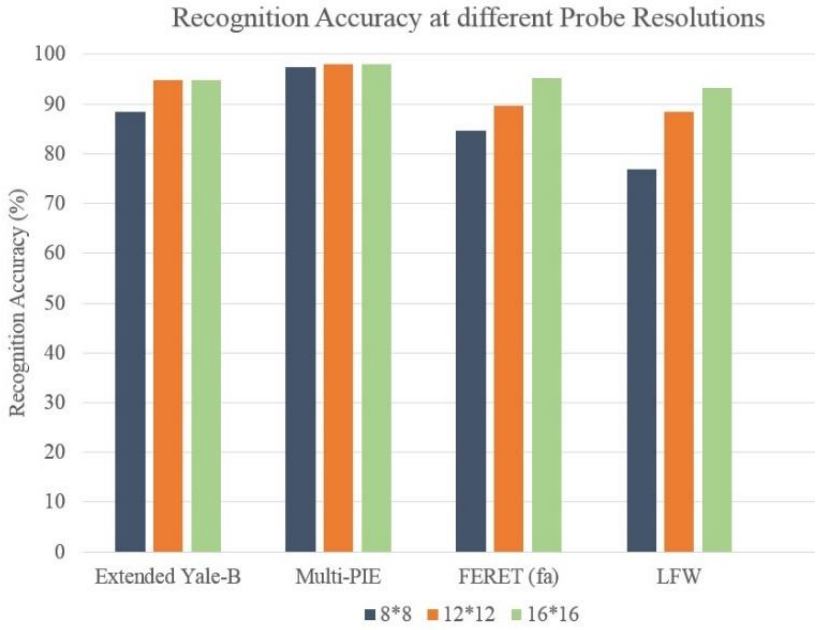


Fig. 3-16. Recognition rates of the proposed method based on four datasets at different probe image resolutions.

Table 3-10. Recognition rates, using different numbers of training images per subject, on the Extended Yale-B database (LR:12 × 12).

Training images / subject	Recognition rate
10	0.9382
20	0.9474
30	0.9461

Table 3-11. Recognition rate, using different numbers of training images per subject, on the Multi-PIE database (LR:8 × 8).

Training images / subject	Recognition rate
10	0.9148
15	0.9720
20	0.9741

Table 3-12. Recognition rate, using different numbers of training images per subject, on the FERET (Fa) database (LR:12 × 12).

Training images / subject	Recognition rate
1	0.8966
2	0.9399

Table 3-13. Recognition rate, using different numbers of training images per subject, on the LFW database (LR: 12×12).

Training images / subject	Recognition rate
5	0.6664
10	0.8823

Table 3-14. Recognition rates of our proposed method, with and without using the morphological pre-processing method.

Database	Recognition rate (without pre-processing)	Recognition rate (after pre-processing)
Extended Yale-B (160-D features)	0.882	**0.947**
Multi-PIE (160-D features)	0.922	**0.974**
FERET (Fa) (200-D features)	0.831	**0.896**
FERET (BaBe) (100-D features)	0.873	**0.915**
LFW (200-D features)	0.819	**0.883**

3.8 Conclusions

This chapter addresses the problem of low-resolution face recognition by proposing a sparse-coding-based approach, which first extracts multiple local features (Gabor wavelets and LBPD) of face images and then decomposes them into a corresponding low-rank feature matrix and a sparse error matrix. The learned low-rank features are then projected into a new discriminative feature subspace using the proposed sparse-coding-based algorithm. It can be observed that sparsity plays an important role in discriminating face images of two different classes. Our proposed method performs sparse reconstruction in the training process, without the need to search any nearest neighbors. The learned projection matrix also preserves the global structure of the data samples in the learned sparse feature subspace. For matching, a coefficient vector is computed to find the similarity between the training and testing image's features by using linear regression. Residual values are then computed based on the estimated coefficient vectors, which represent a testing feature, in terms of a set of training features. Finally, the LR query face image is then assigned to the class label with the least residual value. Our objective function does not need to tune any model parameter. Furthermore, our proposed morphological pre-processing method brings significant improvements in recognition accuracy. Experimental results demonstrate that our proposed method shows excellent performance in recognizing the very LR images even of the size 8×8, which is better than conventional as well as deep-learning-based methods.

Chapter 4 Deep-Feature Encoding-based Discriminative Model for Age-invariant Face Recognition

4.1 Introduction

Face recognition under unconstrained environments has passed various milestones due to the development of various state-of-the-art techniques. However, recognizing face images with aging variations is a challenging research problem that needs considerable amount of attention. It has many practical applications, e.g. criminal identification using photographs, passport verification, etc. Due to age progression, face images go through a major change in terms of both shape and texture, as shown in Fig.4-1.

In the past, various global and local feature descriptors have been proposed to perform face-recognition. As we discussed in the previous chapters, local features have been proven to be robust to various facial variations, such as illumination, expression, pose, etc. However, these features are not optimal for solving cross-age face-recognition problems [178], and provide limited performance. Furthermore, their performance heavily depends on the properly pre-processed face images. To make these features more discriminative for recognition tasks, various feature-encoding-based methods [118-125] were proposed, which convert extracted features into a discriminative codeword for image representation. This brings major improvements in the recognition performance under unconstrained environments. Although deep learning models [48-54] have achieved outstanding performances in solving face recognition problem, but their performance is limited in solving the aging face-recognition problem.

To overcome this problem, we propose a robust deep-feature-encoding method based on locality constraint, which converts extracted deep features into an *N*-dimensional codeword for face representation. In this regard, we first learn an age-discriminative codebook, which ensures the same codeword for the same identity at different ages. Using a pair of face images with a large age difference, we first exploit their correlation by projecting them into a coherent feature subspace using canonical correlation analysis (CCA), and then perform feature fusion. Those fused features are then used to learn an age-discriminative codebook. In the testing stage, the gallery and query image's features are encoded using the learned codebook. The resultant encoded features are sparse, which further improves the discriminative power of the learned features. In the final stage, the linear-regression model [43] is utilized to determine the identity of a given

encoded query feature, in terms of a coefficient vector. By using this coefficient vector, residual values are computed for face matching.

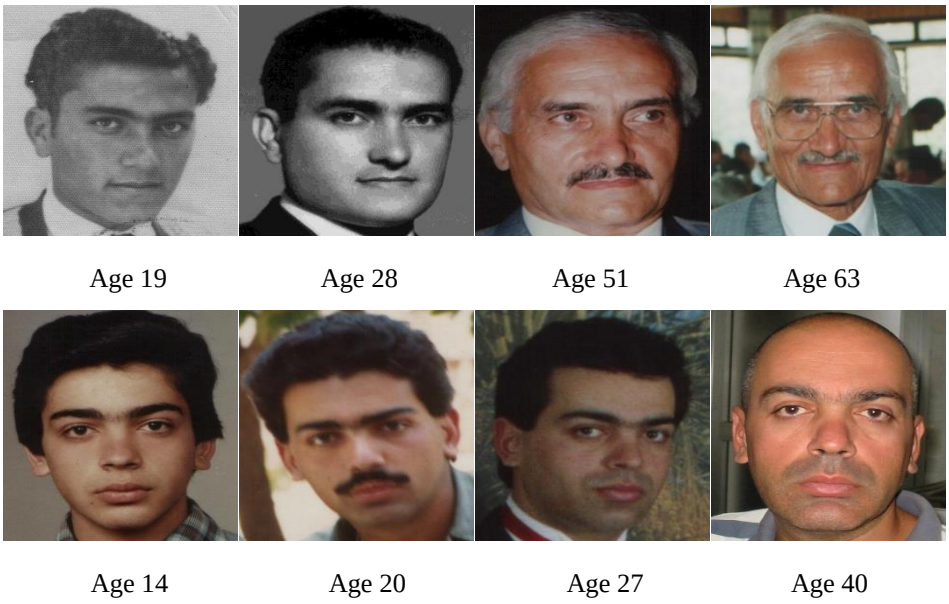


Fig. 4-1. Sample images from the FGNET dataset from two different persons with large age variations, where each row represents the face images of the same person.

4.1.1 Motivation

In terms of face recognition with aging progression, our work is highly inspired by some recent works on age-invariant face recognition (AIFR) [101, 105], which encode pixel values into discriminative codes, such that face images of the same identity at different ages are represented by the same codeword or similar codewords. However, these methods rely on hand-crafted feature descriptors, and do not consider the face images' local structures while performing feature encoding. Data locality is considered as a key issue in many areas of pattern recognition, such as dimensionality reduction, clustering, image classification, etc. Wang et al. [122] proposed a feature-encoding framework based on locality constraint, namely locality constrained linear coding (LLC), which projects extracted features into a local coordinate system. It was claimed that locality information can lead to the sparsity for the resultant encoded coefficients, but not vice versa. Motivated by this idea, we learn an age-discriminative codebook by keeping the data's local structures into account. The locality constraint captures the correlation between the features of the same identity by sharing the local bases of the codebook. Therefore, it ensures the same codeword for the images of the same identity taken at different ages. For cross-age face recognition, these codewords are explored. Furthermore, the utilization of locality information results in the sparse nature of the encoded features, which further enhances the

discriminability of the extracted features. Extracted features of the same person at different ages should have a certain amount of correlation, which can be exploited to learn age-invariant representation. Inspired by the application of CCA, we learn the coherent feature subspace using pairs of the training image's features. Feature fusion is then performed in the learned CCA subspace, which is then used to learn an age-discriminative codebook.

The major contributions of this chapter are as follows:

- A robust feature-encoding framework based on locality constraint is proposed, which encodes the extracted deep facial features into a discriminative codeword. In comparison to LLC [122], our algorithm provides closed form solutions for both encoding and the codebook updating stage.
- In the training stage, we maximize the correlation between the deep features of the same identity with a large age difference using CCA, which are then used to learn an age-discriminative codebook. The learned codebook is proved to be discriminative in terms of age progression, as verified by our experimental results.
- Extensive experiments have been conducted on three challenging face-aging data sets, and experiment results show that our proposed method is capable of recognizing face images with large age gaps, and also outperforms other state-of-the-art AIFR methods, in terms of recognition rate. Furthermore, our proposed method shows robustness to externally added noise, and achieves state-of-the-art performance, as verified by our experimental results.

The remaining chapter is organized as follows. In Section 4.2, we introduce the process of deep-feature extraction using a deep-CNN model (AlexNet). In Section 4.3, we present our feature-encoding framework based on the Euclidean locality adaptor, which makes the extracted deep features more robust to aging variations. In Section 4.4, we explain the concept of feature fusion using CCA. In Section 4.5, we discuss the use of linear regression for classification. Section 4.6 presents the experimental results of our proposed discriminative model, along with discussions. Finally, the chapter is concluded in Section 4.7.

Our proposed discriminative model consists of four main parts: (1) deep-feature extraction, (2) age-discriminative codebook learning, (3) feature-encoding using learned codebook, and (4) face matching based on linear mapping. We will explain all the stages in the following sections. After that, we will present the comprehensive analysis of our experimental results.

4.2 Deep-Feature Extraction

In our work, we utilize a pre-trained deep-CNN model, namely AlexNet [179], to extract high-level features for AIFR. CNN models have the capability of learning effective features from input images. AlexNet is selected, due to its simple architecture and superior performances. Fig. 4-2 shows the AlexNet architecture. It consists of five convolutional layers, three pooling layers, and three fully connected (FC) layers. The output of each convolutional and fully connected layer is fed to the ReLU. The first convolutional layer filters the input face image of size $227 \times 227 \times 3$, using 96 kernels of size $11 \times 11 \times 3$. The output of the first convolutional layer is fed to the second convolutional layer, after passing through the normalization and pooling layer. After passing through the five convolutional layers, the network learns high-level deep features at the fully connected layer, with a dimension of 4096. All fully connected layers are regularized by using a drop-out scheme. In our method, we extract features from the FC layer ('fc7').

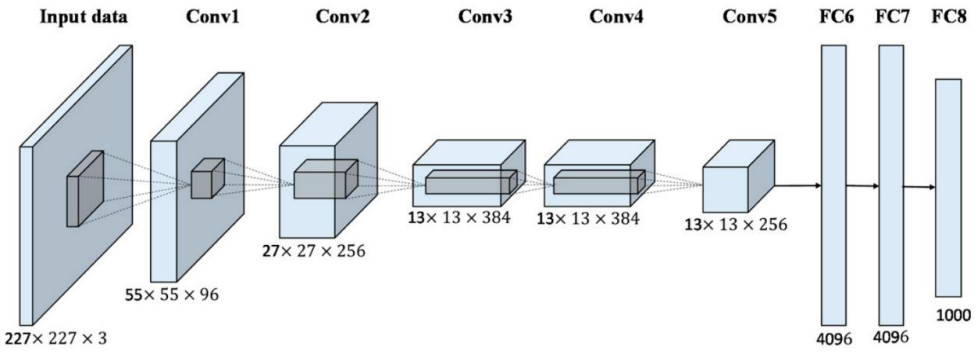


Fig. 4-2. AlexNet architecture (Adapted from paper [197]).

In deep learning-based methods, networks are trained for feature extraction and recognition from end to end. In our proposed approach, we use deep-learning-based CNN model to extract features, which are then converted into discriminative codewords for recognition. Fig. 4-3 shows the features learned at the different layers of the CNN.

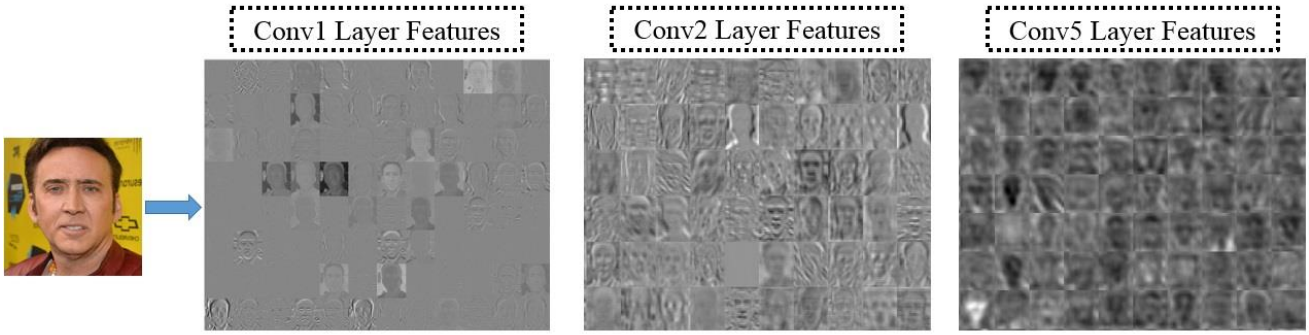


Fig. 4-3. Visualization of the learned deep features from different convolutional layers of AlexNet.

4.3 Feature Encoding based on locality information

We now explain our proposed feature-encoding framework, based on the Euclidean locality adaptor. Before we proceed, we first explain the major differences between data sparsity and data locality. The advantages of using locality information for accurate classification (recognition) will also be discussed. Our proposed method provides a better feature-encoded representation, which enhances the discriminability of the extracted deep features in terms of age progression.

4.3.1 Locality vs Sparsity

Data locality information has been proven to be important for the success of various pattern-recognition applications, e.g. density estimation [180], dimensionality reduction [30], and image classification [181]. SRC [7] was used for face recognition under large occlusion, and achieved state-of-the-art performances. However, is it enough, or essential, for resolving this kind of a problem? This query has recently been examined in [122], which concluded that exploitation of sparse information only is not enough to handle pattern recognition problems with large occlusion.

As discussed before, the data's locality has more importance than sparsity, as sparse encoded coefficients can be achieved using locality, but not vice versa. By using locality information, codebook entries near to the query input will be selected for the reconstruction of data samples. However, classification based on sparse representation minimizes the class-wise reconstruction error, so using codebook entries far away from the query input for reconstruction is not desirable. According to the assumption made by the *k*NN classifier, those codebook entries far away from the input have less probability of belonging to the same class. The standard feature-encoding method [120], based on data sparsity, do not incorporate the local structure information of the data samples during the encoding stage, but LLC [122] does. This encourages us to propose a feature-encoding-based framework based on the Euclidean locality adaptor, which preserves the data structure and has better feature representation, and hence, enhances the recognition performance.

4.3.2 Locality-based feature-encoding framework

Now, we present our feature-encoding framework based on the Euclidean locality adaptor, which makes the extracted deep features discriminative in terms of age progression. We propose to encode the features by projecting each of them into its local coordinate system. Consider that M feature vectors of dimension D are extracted, and they are represented as $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_M] \in R^{D \times M}$. Firstly, a codebook $\boldsymbol{W}$ with N entries, i.e. $\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_N\}$ is generated using

k-means clustering, which is then used to convert each deep feature into a N-dimensional codeword for final image description. Our proposed feature-encoding approach focuses on the Euclidean locality constraint rather than the sparsity constraint. The proposed locality-based objective function is defined as follows:

$$\min_{\boldsymbol{W},\boldsymbol{C}} \|\boldsymbol{F} - \boldsymbol{W}\boldsymbol{C}\|^2 + \lambda \sum_{k=1}^{M} \|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2, \tag{4.1}$$

$$\text{s.t. } \mathbf{1}^T \boldsymbol{c}_k = 1,$$

where $\otimes$ represents the element-wise multiplication operator, λ is a regularization parameter, $\boldsymbol{l}_k \in R^N$ represents the locality term that consists of a Euclidean adaptor, which provides freedom to each basis vector depending on how similar it is to the given descriptor $\boldsymbol{f}_k$, and $\mathbf{1}$ is the identity vector $[1, \dots, 1]^T$. $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_M]$ is the set of codes for $\boldsymbol{F}$. Each entry of the locality term $\boldsymbol{l}_k$ can be defined as follows:

$$\boldsymbol{l}_{kn} = (\sigma^2 + \|\boldsymbol{f}_k - \boldsymbol{w}_n\|^2)^{-1}\ , \tag{4.2}$$

where $\|\boldsymbol{f}_k - \boldsymbol{w}_n\|^2$ is the Euclidean distance between the input feature descriptor $\boldsymbol{f}_k$ and the n^{th} codebook entry $\boldsymbol{w}_n$. σ is a constant that controls the weight decay speed for the locality term. In our method, we choose $\sigma = 0.5$. Later on, we will justify the selection of this constant value. The locality term in equation (4.2) is the Euclidean adaptor that defines how local coding varies with respect to the distances $\|\boldsymbol{f}_k - \boldsymbol{w}_n\|$. It utilizes the student t-distribution to provide a degree of freedom, such as Cauchy distribution. One of the major properties of student t-distribution is that $(\sigma^2 + \|\boldsymbol{f}_k - \boldsymbol{w}_n\|^2)^{-1}$ follows an inverse square law when the pairwise distance $\|\boldsymbol{f}_k - \boldsymbol{w}_n\|^2$ is large. $\boldsymbol{l}_k$ is further normalized to have a value between (0, 1) by taking a difference between $\max(\|\boldsymbol{f}_k - \boldsymbol{w}_n\|^2)$ and $\|\boldsymbol{f}_k - \boldsymbol{w}_n\|^2$. Furthermore, local bases $\boldsymbol{w}_k$ are selected for each feature descriptor, such that a local coordinate system can be built. Local bases can be considered as the nearest neighbours of $\boldsymbol{f}_k$, which lead to a more compact and simplified linear system for feature coding.

To solve (4.1), we utilize the Lagrange multiplier, which is defined as follows:

$$L(\mathbf{c}_k, \eta) = \|\boldsymbol{f}_k - \boldsymbol{W}\boldsymbol{c}_k\|^2 + \lambda\|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2 + \eta(\mathbf{1}^T\boldsymbol{c}_k - 1). \tag{4.3}$$

Let $\mathbf{Y} = (\boldsymbol{f}_k\mathbf{1}^T - \boldsymbol{W})^T(\boldsymbol{f}_k\mathbf{1}^T - \boldsymbol{W})$, which is symmetrical. Equation (4.3) can be rewritten as:

$$L(\boldsymbol{c}_k, \eta) = \boldsymbol{c}_k{}^T\boldsymbol{Y}\boldsymbol{c}_k + \lambda\boldsymbol{c}_k{}^T\text{diag}(\boldsymbol{l}_k)^2\boldsymbol{c}_k + \eta\left(\mathbf{1}^{\mathrm{T}}\boldsymbol{c}_k - 1\right), \tag{4.4}$$

where $\mathrm{diag}(\boldsymbol{l}_k)$ is a diagonal matrix. To determine the optimal solution of (4.4), its partial derivative is set to zero, which gives the following equation:

$$\frac{\partial(L(\mathbf{c}_k,\eta))}{\partial(\mathbf{c}_k)} = 2\boldsymbol{Y}\boldsymbol{c}_k + 2\lambda\mathrm{diag}(\boldsymbol{l}_k)^2\boldsymbol{c}_k + \eta^{\mathrm{T}}\mathbf{1} = 0\,. \tag{4.5}$$

Let $\boldsymbol{\Phi} = 2(\boldsymbol{Y} + \lambda\mathrm{diag}(\boldsymbol{l}_k)^2)$, we have

$$\boldsymbol{\Phi}\boldsymbol{c}_k + \eta\mathbf{1} = 0\ , \tag{4.6}$$

Multiply (4.6) by $\mathbf{1}^T\Phi^{-1}$, the following equation is obtained:

$$\mathbf{1}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}\boldsymbol{c}_k + \eta(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1}) = 0 \tag{4.7}$$

According to the constraint $\mathbf{1}^T\boldsymbol{c}_k = 1$, $\mathbf{1}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}\boldsymbol{c}_k = 1$, which gives us:

$$1 + \eta(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1}) = 0,$$

$$\eta = -(\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1})^{-1}. \tag{4.8}$$

Putting η into (4.6), we obtain the following equation:

$$\boldsymbol{\Phi}\boldsymbol{c}_k = (\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1})^{-1}\mathbf{1}.$$

After doing some transformations, we have

$$\boldsymbol{c}_k = \frac{\boldsymbol{\Phi}^{-1}\mathbf{1}}{\mathbf{1}^T\boldsymbol{\Phi}^{-1}\mathbf{1}} = \frac{\frac{1}{2}\boldsymbol{\Phi}^{-1}\mathbf{1}}{\mathbf{1}^T\left(\frac{1}{2}\boldsymbol{\Phi}^{-1}\mathbf{1}\right)}. \tag{4.9}$$

In this way, we obtain the analytical solution of our proposed objective function, which is given as follows:

$$\tilde{\boldsymbol{c}}_k = \tfrac{1}{2}\boldsymbol{\Phi}^{-1}\mathbf{1} = (\boldsymbol{Y} + \lambda\mathrm{diag}(\boldsymbol{l}_k)^2)^{-1}\mathbf{1},$$

$$\boldsymbol{c}_k = \tilde{\mathbf{c}}_k/(\mathbf{1}^T\tilde{\boldsymbol{c}}_k). \tag{4.10}$$

To update the codebook $\boldsymbol{W}$, it needs to solve the following equation:

$$\min_{\boldsymbol{W}}\|\boldsymbol{F} - \boldsymbol{W}\boldsymbol{C}\|^2 + \lambda\sum_{k=1}^{M}\|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2. \tag{4.11}$$

Let the objective function (4.11) be denoted as $\boldsymbol{F}(\boldsymbol{W})$, which update the codebook initialized using *k*-means clustering. Analytical solution of (4.11) can be derived by taking the partial derivative of $\boldsymbol{F}(\boldsymbol{W})$ with respect to $\boldsymbol{w}_n$ for $n \in \{1,2,\dots,N\}$, which gives us the following equation:

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{w}_n} = \sum_{k=1}^{M} -2c_{kn}(\boldsymbol{f}_k - \boldsymbol{W}\boldsymbol{c}_k) - 2\lambda c_{kn}^2(\boldsymbol{f}_k - \boldsymbol{w}_n) \tag{4.12}$$

In an equivalent form, it can be written as:

$$\left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{w}_n}\right)^T = \sum_{k=1}^{M}(-2c_{kn}(1+\lambda c_{kn})(\boldsymbol{f}_k)^T + 2(\lambda c_{kn}^2 \boldsymbol{w}_n{}^T + c_{kn}\sum_{j=1}^{N} c_{kj}\boldsymbol{w}_j{}^T)) \,. \tag{4.13}$$

To compute the global minimum of (4.11), we set its partial derivatives to zero. After setting the partial derivative of (4.11) to zero for $n = 1, 2, 3, \dots, N$, we obtain

$$\boldsymbol{P}\boldsymbol{W}^T = \boldsymbol{Q}, \tag{4.14}$$

where the matrices $\boldsymbol{P} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$ are

$$\boldsymbol{P} = \sum_{k=1}^{M}\begin{pmatrix} (1+\lambda)c_{k1}^2 & c_{k1}c_{k2} & \cdots & c_{k1}c_{kN} \\ c_{k1}c_{k2} & (1+\lambda)c_{k2}^2 & \cdots & c_{k2}c_{kN} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1}c_{kN} & c_{k2}c_{kN} & \cdots & (1+\lambda)c_{kN}^2 \end{pmatrix},$$

$$\boldsymbol{Q} = \sum_{k=1}^{M}\begin{pmatrix} c_{k1}(1+\lambda c_{k1})(\boldsymbol{f}_k)^T \\ c_{k2}(1+\lambda c_{k2})(\boldsymbol{f}_k)^T \\ \vdots \\ c_{kN}(1+\lambda c_{kN})(\boldsymbol{f}_k)^T \end{pmatrix}. \tag{4.15}$$

Finally, the updated codebook can be obtained by solving the equation (4.14). Therefore, the optimal solutions for encoding the parameters $\boldsymbol{C}$ and the codebook $\boldsymbol{W}$ are obtained, using coordinate descent method. In this process, we optimize $\boldsymbol{C}(\boldsymbol{W})$ based on the existing value of $\boldsymbol{W}(\boldsymbol{C})$, alternately. The solutions of both $\boldsymbol{C}$ and $\boldsymbol{W}$ are unique, and their sequences also converge to stationary points. In the encoding stage, when codebook $\boldsymbol{W}$ is fixed, we derive the analytical solution of $\boldsymbol{C}$ using (4.10). Similarly, during the codebook-updating process, the closed form solution of $\boldsymbol{W}$ is derived using (4.14). During the testing stage, each query and gallery face image's feature is encoded using the learned codebook $\boldsymbol{W}$. Our experiments also show that, by encoding features at different ages using locality constraint, we can

obtain more discriminative feature representation for age-invariant face recognition. Algorithm 4-1, and 4-2 summarizes the training and testing steps of our proposed feature-encoding framework.

Algorithm 4-1: Codebook updating scheme

Input: $\boldsymbol{W}_{init} \in R^{D\times N}, \boldsymbol{F} \in R^{D\times M}, \sigma, \lambda$

Output: $\boldsymbol{W}$

1: $\boldsymbol{W} \leftarrow \boldsymbol{W}_{init}$ (Initialize the codebook using *k*-means)

2: **for** $k = 1: M$ **do**

3: $\boldsymbol{l} \leftarrow 1 \times N$ (Locality constraint)

4: **for** $k = 1: N$ **do**

5: $\boldsymbol{l}_{kn} \leftarrow (\sigma^2 + \|\boldsymbol{f}_k - \boldsymbol{w}_n\|^2)$

6: end **for**

7**:** $\boldsymbol{l} \rightarrow normalize_{(0,1)}(\boldsymbol{l})$

8**:** Lagrange function to solve (4.1)

$L(\boldsymbol{c}_k, \eta) = \|\boldsymbol{f}_k - \boldsymbol{W}\boldsymbol{c}_k\|^2 + \lambda\|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2 + \eta(\boldsymbol{1}^T\boldsymbol{c}_k - 1)$

$\boldsymbol{c}_k = \tilde{\boldsymbol{c}}_k/(\boldsymbol{1}^{\mathrm{T}}\tilde{\boldsymbol{c}}_k)$ Eq. (4.3) − (4.10) (Analytical solution)

9: Codebook Updating:

$$\left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{w}_n}\right)^T = \sum_{k=1}^{M}(-2c_{kn}(1 + \lambda c_{kn})(\boldsymbol{f}_k)^T + 2(\lambda c_{kn}^2 \boldsymbol{w}_n{}^T + c_{ks}\sum_{j=1}^{N} c_{kj}\boldsymbol{w}_j{}^T))$$

10: $\boldsymbol{P}\boldsymbol{W}^T = \boldsymbol{Q}$

11: end **for**

12**:** Updated Codebook $\boldsymbol{W}$

4.4 Feature fusion using CCA

Codebook learning is the most important and critical step of our proposed algorithm. To recognize faces across different ages, the learned codebook must be discriminative in terms of age progression. To do this, we perform feature-level fusion using CCA. The simplest way of doing feature fusion is by computing the z-score, which is done by normalizing the two feature vectors and then concatenating them to create a high-dimensional feature vector. However, it does not take the correlation between the two features into account. In our algorithm, features are extracted and encoded at different ages. Features of the same person at different ages should have a certain degree of correlation.

Therefore, we first project the features of the training image's pairs into a coherent feature subspace and then concatenate them to form the final feature vector. This concatenated feature vector is used to learn a discriminative codebook. Given two sets of training features, with each pair extracted from images of the same subject but at different ages, denoted as $\boldsymbol{F}_{age1}$, and $\boldsymbol{F}_{age2}$, we employ CCA to learn the pairs of directions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which maximize the correlation among the two aging features, such that $q_j^1 = \boldsymbol{\alpha}^T \boldsymbol{F}_{age1}$ and $q_j^2 = \boldsymbol{\beta}^T \boldsymbol{F}_{age2}$, where $\boldsymbol{F}_{age1}$ and $\boldsymbol{F}_{age2}$ are the features extracted from a younger and an older images, respectively. q_j^1 and q_j^2 are the projected training image's features. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be computed by maximizing the following function.

$$K(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{C}_{xy} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{C}_{xx} \boldsymbol{\alpha}. \boldsymbol{\beta}^T \boldsymbol{C}_{yy} \boldsymbol{\beta}}}. \tag{4.16}$$

where $\boldsymbol{C}_{xx}$ and $\boldsymbol{C}_{yy}$ are the covariance matrices of ${q_j}^1$ and ${q_j}^2$, respectively, while $\boldsymbol{C}_{xy}$ is the cross-variance matrix. After solving, it can be observed that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the eigenvectors of ${\boldsymbol{C}_{xx}}^{-1} \boldsymbol{C}_{xy} {\boldsymbol{C}_{yy}}^{-1} {\boldsymbol{C}_{xy}}^T$, and ${\boldsymbol{C}_{yy}}^{-1} {\boldsymbol{C}_{xy}}^T {\boldsymbol{C}_{xx}}^{-1} \boldsymbol{C}_{xy}$, respectively. Fig. 4-4, and 4-5 shows the training and the testing stage of our proposed framework.

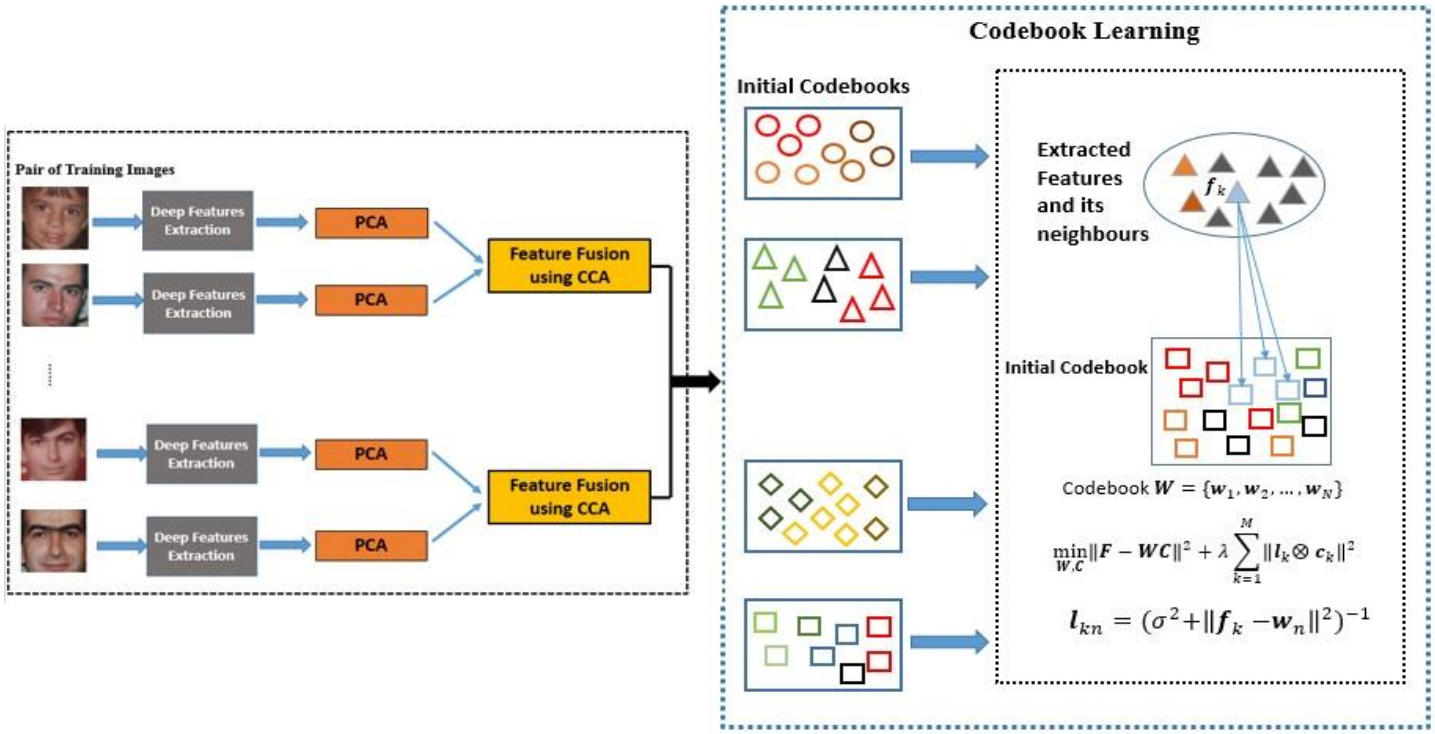


Fig. 4-4. Training Stage of our proposed framework.

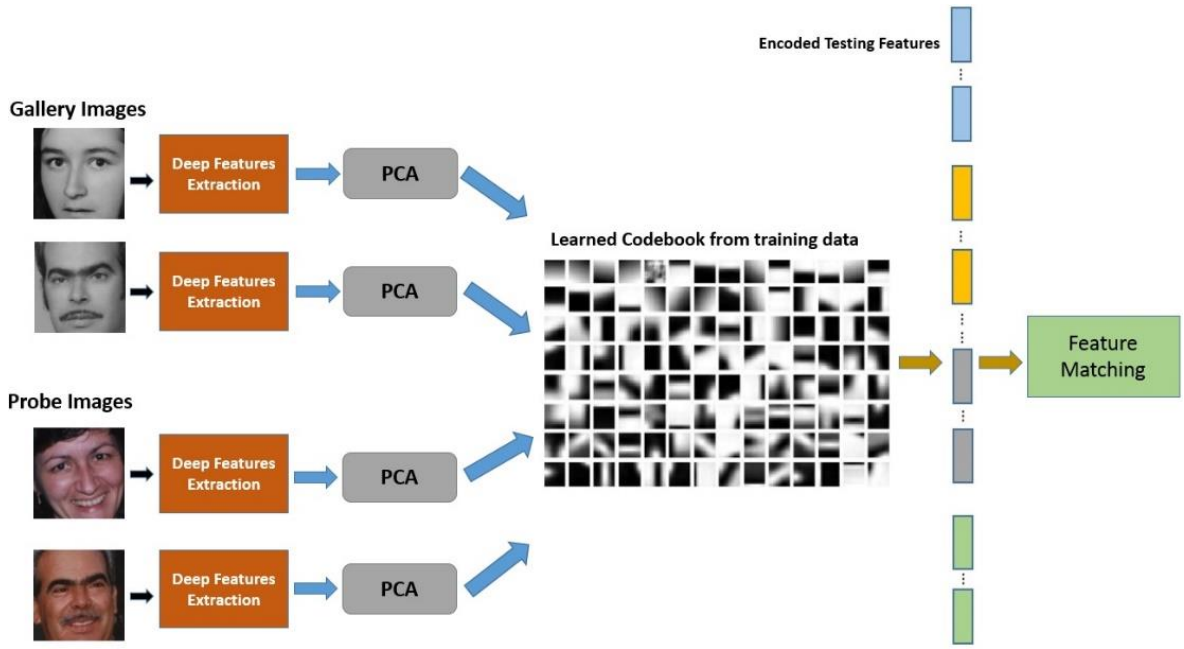


Fig. 4-5. Testing stage of our proposed framework.

Algorithm 4-2: Feature encoding using a learned Codebook (Testing Stage)

1: **Input**: $\boldsymbol{W}_n$ for $n \in \{1,2,\dots,N\}$, $\boldsymbol{q}$ be a feature of a query image.

2: for $k = 1 : N_{knn}$ **do** (where N_{knn} is number of nearest neighbors (number of local bases selected from the codebook)

3: First Euclidean locality adaptor $\boldsymbol{l}_k$ is computed

$$\boldsymbol{l}_{kn} = (\sigma^2 + \|\boldsymbol{q} - \boldsymbol{w}_{kn}\|^2)^{-1}$$

4: Solve the Equation (4.1) to obtain the analytical solution

$$\min_{\boldsymbol{c_k}} \|\boldsymbol{q} - \boldsymbol{W_k c_k}\|^2 + \lambda \|\boldsymbol{l_k} \otimes \boldsymbol{c}_k\|_2^2$$

$$\boldsymbol{\beta}^k = (\boldsymbol{Y}_k + \lambda \text{diag}(\boldsymbol{l}_k)^2)^{-1}\mathbf{1}$$

$$\boldsymbol{c}_k = \boldsymbol{\beta}^k / (1^T \boldsymbol{\beta}^k)$$

where $\boldsymbol{Y}_k = (\boldsymbol{q}\mathbf{1}^T - \mathbf{W}_k)^T (\boldsymbol{q}\mathbf{1}^T - \boldsymbol{W}_k)$

5: Similarly, encoding gallery image's features using step (4). Let the set of codewords obtained for all Gallery features is denoted as $\boldsymbol{G}_k$.

6: Compute coefficient vector $\boldsymbol{\gamma}_k$ using Equation (4.17).

7: Compute the residuals $r_k(\boldsymbol{q}) = \|\boldsymbol{c}_k - \boldsymbol{G}_k \boldsymbol{\gamma}_k\|$

8: end **for**

9: identity$(\boldsymbol{q}) = \arg\min_k r_k(\boldsymbol{q})$

(a)

(b)

Fig. 4-6. Sample images with age variations, where each row represents the face images of the same person. (a) Morph dataset, and (b) LAG dataset.

4.5 Feature Matching Using Linear Regression

In order to determine the similarity between the encoded gallery and the query images' features, we utilize a linear regression model [43]. The assumption is that the features of different samples lie in a linear subspace, so a query face image can be linearly represented in terms of all face images in the gallery. The relationship between the query (test) and training sample is determined by a parameter $\boldsymbol{\gamma}$, known as coefficient vector. This parameter is estimated by using the least-squares method. It is also considered as a prediction problem having a solution based on a regression framework. If a query image $\boldsymbol{q}$ belongs to the k^{th} class, then there must exist a linear relationship between this query image and the gallery samples $\boldsymbol{X}_k$ from the same class, which is defined as follows:

$$\boldsymbol{q} = \boldsymbol{X}_k \boldsymbol{\gamma}_k \,, \tag{4.17}$$

After estimating the coefficient vector $\boldsymbol{\gamma}$, the corresponding residual values are computed. The decision will be in favor of the query image, if it has a minimum distance to the gallery image. It can be written as:

$$j = \min_k \|\boldsymbol{q} - \boldsymbol{X}_k \boldsymbol{\gamma}_k\|. \tag{4.18}$$

4.6 Experimental Results and Analysis

Creating a data set with a large age variation is a difficult task. Currently, only a few aging datasets are available, which limits the research on age-invariant face recognition. For comprehensive analysis of our proposed AIFR algorithm, face datasets should have the following attributes: (1) a large number of samples per subject, (2) significant age variations among the images of the same person, and (3) images must be taken in unconstrained environments. The execution of our proposed method is assessed by conducting extensive set of experiments on three challenging face-aging data sets: FGNET [144], MORPH [182], and LAG datasets [183]. Images in FGNET and LAG datasets are taken in the wild, and have large age gaps. For all the three data sets, we first detect the location of the face region in an image using the Viola-Jones face detector [17], and then resize the face region to 227×227 pixels. In addition to linear regression (LR) based classifier, we also utilize nearest neighbor (NN) classifier for feature matching. Moreover, the performance of our proposed feature-encoding framework is also evaluated by fusing two efficient local features, namely Dense SIFT (DSIFT) [39] and local binary pattern difference feature (LBPD) [153].

4.6.1 Experimental Results on the FGNET Database

FGNET is considered as one of the most challenging face aging datasets, which contains 1,002 images of 82 subjects taken at different ages. The minimum age of a person in this dataset is less than 12 months, and maximum being 69. As the number of subjects are quite small, so more images per subject are available. In addition to aging variations, images in this data set also contain large variations in terms of expression, illumination, and pose. By following the same protocol as used in [14, 102], we used leave-one-out cross-validation (LOOCV) scheme for performance evaluation. Furthermore, our experimental results are compared with various state-of-the-art AIFR methods. These include: (1) the 3D age modeling technique for age estimation and recognition [91]; (2) a discriminative model for AIFR [14]; (3) hidden factor analysis [102], which represents face images with the identity and age components; (4) the maximum entropy feature descriptor [101], and (5) deep-learning approach based on latent factor guided convolutional neural networks [106], which has achieved the highest recognition rate so far on this most challenging face aging data set, and (6) method based on coupled auto-encoder network [109]. From Table 4-1, we can observe that our proposed feature-encoding-based discriminative model clearly outperforms other AIFR methods, and achieves the highest rank-1 recognition rate.

Table 4-1. Comparative results in terms of the Rank-1 recognition rate on the FGNET dataset.

Algorithms	Rank-1 Recognition rates
Park et al. [91]	37.4%
Li et al. [14]	47.5%
Gong et al. [102]	69.0%
MEFA [101]	76.2%
CNN-baseline	84.4%
LF-CNN [106]	88.1%
Xu et al. [109]	86.5%
Proposed Method (DSIFT+LBPD) + NN	**90.48%**
Proposed Method (DSIFT+LBPD) + LR	**89.13%**
Proposed Method (Deep features) + NN	**91.46%**
Proposed Method (Deep features) + LR	**90.24%**

Table 4-2. Comparative results in terms of the Rank-1 recognition rate on the MORPH database (Album 2).

Algorithms	Rank-1 Recognition rate
Park et al. [91]	79.80%
Li et al. [14]	83.90%
Gong et al. [102]	91.14%
MEFA [101]	92.26%
CARC [103]	92.80%
HOG+LPS [105]	94.20%
LPS+HFA	94.87%
LF-CNN [106]	97.51%
AFJT-CNN [108]	97.85%
Proposed Method (DSIFT+LBPD) + NN	**96.06%**
Proposed Method (DSIFT+LBPD) + LR	**96.50%**
Proposed Method (Deep features) + NN	**97.93%**
Proposed Method (Deep features) + LR	**98.00%**

4.6.2 Experimental Results on the MORPH Database

We also conducted experiments on one of the largest face-aging data set, i.e. MORPH Album 2. This data set contains 78,000 face images from 20,000 subjects. The number of images per subject are small (around 4 images per person) due to the availability of a large number of subjects. Firstly, the dataset is split into a training and a testing set. To learn a codebook for feature encoding, 20,000 images from 10,000 different subjects (2 images per person), with the large age difference are used. For testing, a gallery set and a probe set are constructed from the other 10,000

subjects; the probe set consists of 10,000 face images of the oldest age of the subject, while the gallery set consists of 10,000 face images of the youngest age. The age gap in this data set is 5-6 years. It should be noted that the age difference between the gallery set and a probe set is quite large. We compared our method with conventional as well as deep-learning-based AIFR methods. Comparative results are tabulated in Table 4-2. Results show that our method significantly outperforms other AIFR methods in comparison, and achieves the highest recognition rate. Sample face images from this dataset are shown in Fig.4-6 (a).

4.6.3 Experimental Results on the LAG Data Set

The Large Age-Gap (LAG) data set [183] is recently released for studying the cross-age face-recognition problem. All the images were taken in the wild, with a very large age difference (0-80) yrs. The dataset is created using a Google search. It consists of 3,828 images from 1,010 identities. At least one child and one adult image is included for each identity. Sample face images from the LAG data set are shown in Fig. 4-6(b). For performance evaluation, we utilize a two-fold cross validation scheme. Subjects are alternatively assigned to the first and the second fold, and then the average accuracy is computed. The training set consists of the original images and a combination of four horizontal flips from the LAG data set. We compare the results of our proposed method with the state-of-the-art high-dimensional LBP feature [127], and some similarity metric learning methods, including Cosine similarity [184], one shot similarity kernel (OSS) [185], and Joint Bayesian [186], and sub-SML [187]. It should be noted that all these similarity metric learning methods used the $fc7$ features extracted from DCNN [3] (trained on the CASIA-WebFace data set [57]). Comparative results are shown in Table 4-3. Some of the correct matching results are shown in Fig. 4-7.

Table 4-3. Comparative results in terms of Rank-1 average recognition rates, on the LAG database.

Algorithms	Rank-1 recognition rate
DCNN [3]+SML [187]	72.43%
DCNN [3]+OSS [185]	66.42%
DCNN[3] + Cosine Similarity [184]	65.08%
DCNN [3] + Joint Bayesian [186]	66.33%
DCNN [3] + CARC [103]	74.82%
HDLBP [127]	71.53%
Bianco et al. [183]	84.95%
Proposed Method (DSIFT+LBPD)+NN	79.88%
Proposed Method (DSIFT+LBPD)+LR	80.00%
Proposed Method (Deep features)+NN	**91.00%**
Proposed Method (Deep features)+LR	**89.44%**

(a) (b)

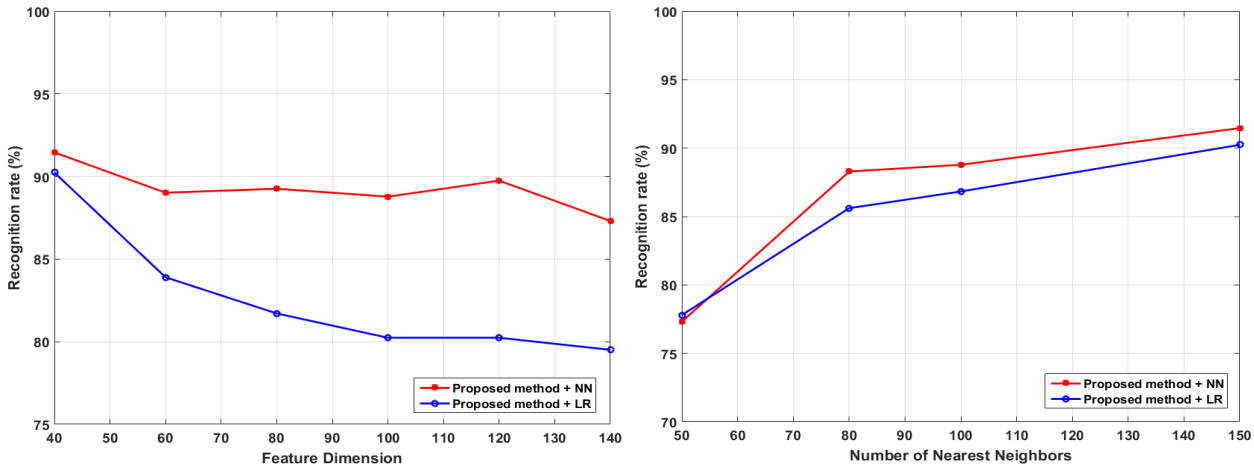
Fig. 4-7. Some of the correct matching results obtained using our proposed method. First columns in (a) and (b) represents the probe images, while the second column represents the identified images from the gallery set.

4.6.4 Parameter Settings

To accelerate the encoding process, some specific number of local bases are selected from codebook to encode the gallery and query image's features, as mentioned in section 4.3. Therefore, we perform comprehensive analysis of our proposed discriminative model, by computing the recognition rate with respect to different numbers of selected nearest neighbors (local bases) of the codebook for encoding. As discussed previously, the codebook is first generated by using *k*-means clustering. As the results on the FGNET data set are evaluated using the LOOCV scheme, we initialize the codebook with $N - 1$ entries, where N is the number of subjects in the data set. The computational complexity of the algorithm depends on the number of nearest neighbors of the feature descriptor $\boldsymbol{f}_k$, which can also be considered as the local bases $\boldsymbol{w}_k$. A small number of neighbors will lead to a faster computation.

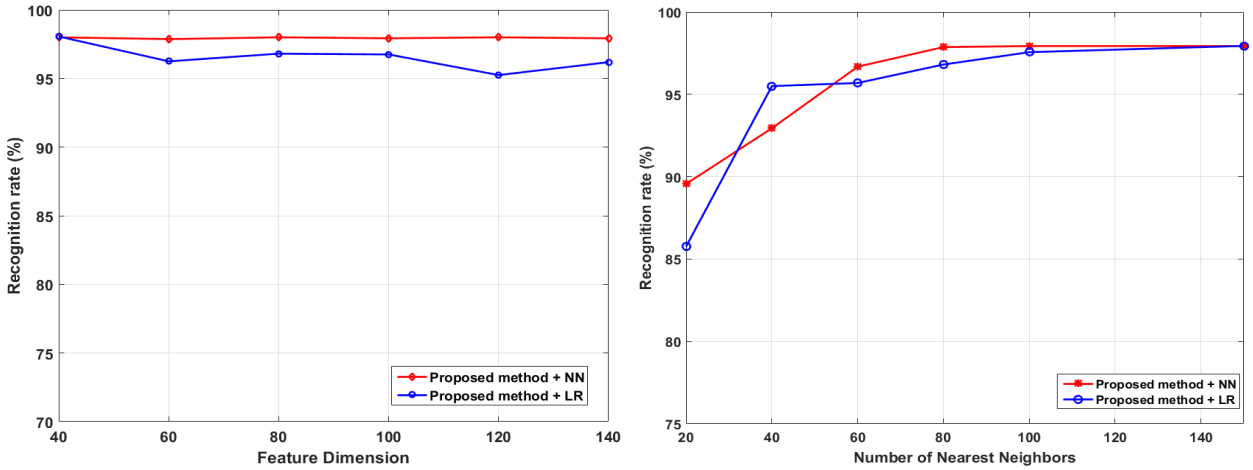
To search the nearest neighbors, the *K*-NN search approach based on the hierarchical model [188] is utilized. The approach quantizes each descriptor into P subspaces. A codebook is then applied for each subspace. For the FGNET data set, we measured the recognition rates with different numbers of nearest neighbors, from 50 to 150, and the results are shown in Fig. 4-8(b). When we increase the number of neighbors, the recognition rate also increases. With the FG-NET data set, the highest recognition rate obtained is 91.46%, by searching for 150 nearest neighbors to form the local bases in the computation of the codes. For the MORPH data set, recognition results are recorded with different numbers of nearest neighbors, ranging from 20 to 150, as shown in Fig. 4-9(b). The highest recognition rate obtained is 98.00%, which is very close to the deep-learning-based method [108]. It is found that the modeling capacity can be greatly improved by using a larger codebook. For the LAG data set, the highest recognition rate of 89.44% is achieved, using the linear-regression (LR)-based classifier. The parameter σ in Equation (4.2) is set to 0.5. The value of σ controls the

locality error of the encoding scheme. Theorem 2 in [189] shows that locality error reduces, as the value of σ decreases. Therefore, the value of σ must be as small as possible. In our experiments, we vary the value of σ from 0.1 to 0.7, and found that $\sigma = 0.5$ gives the optimum performance. The choice of the parameter λ in Equation (4.1) can be well explained by Equation (4.10). It is worth noting that the matrix $\boldsymbol{Y}$ is symmetric as well as semi-positive. If $\boldsymbol{Y}$ becomes singular or close to singular, the matrix $\boldsymbol{Y} + \lambda \text{diag}(\boldsymbol{l}_k)^2$ is still conditioned. The reason for this is that $\lambda \text{diag}(\boldsymbol{l}_k)^2$ penalizes the large distance, which captures the correlation between the data samples. By choosing $\lambda < 10^{-6}$, the matrix becomes singular or close to singular, which will produce inaccurate results. Therefore, choosing $\lambda = 0.001$ provides the optimal recognition results. Moreover, we also perform experiments, using deep features of different dimensions, on all the three datasets, and results are shown in Figs. 4-8, 4-9, and 4-10(a), respectively.



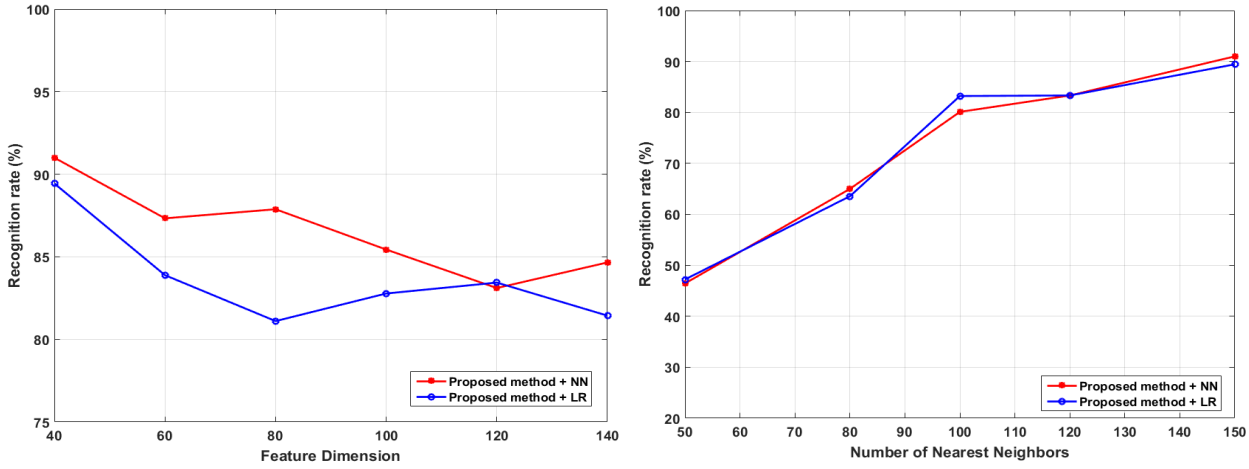
(a) (b)

Fig. 4-8. Recognition rates obtained on the FGNET data set. (a) Feature Dimensions with 150-NN; (b) Number of nearest neighbours (40-D features).



(a) (b)

Fig. 4-9. Recognition rates obtained on the MORPH data set. (a) Feature Dimensions with 150-NN; (b) Number of nearest neighbours (40-D features).



(a) (b)

Fig. 4-10. Recognition rates obtained on the LAG data set. (a) Feature Dimensions with 150-NN; (b) Number of nearest neighbours (40-D features).

To check the effectiveness of our proposed feature encoding framework, we evaluate the performance with and without performing the proposed feature-encoding scheme. The results using the nearest neighbor (NN) classifier and the linear regression (LR) classifier are shown in Fig. 4-11. It can be observed that our proposed feature-encoding framework boosts the recognition accuracy by 20-35%.

As discussed before, CCA enhances the correlation among the images of the same subject taken at different ages. To evaluate the superiority of CCA-based feature fusion, we perform the comparative analysis of CCA-based feature fusion (CCA_FF) and simple concatenation of the two features. Experimental results are reported for all the three datasets at their optimal feature dimensions with different numbers of nearest neighbors used for feature encoding. The results are reported in Fig. 4-12. It can be observed that the feature fusion framework based on CCA brings significant improvement in recognition rate.

Another important parameter of our proposed framework is the number of image pairs per person, used for the CCA pairwise training. For the FGNET dataset, 10 images per subject are available. Therefore, we divide these images into two subsets, with 5 images each, and then use them for CCA training. One subset consists of those younger images, while the other one contains the older images. As the performance on the FGNET data set was evaluated using the LOOCV scheme, we used 81 image pairs for training. For the MORPH data set, 10,000 image pairs were used. The training set consists of 10,000 subjects, with the two images having the largest age difference. Images of each subject were divided equally into two subsets for pairwise CCA training. For the LAG data set, the two-fold scheme was used

for performance evaluation. For training, we selected eight images per subject, with a large age difference. As the total number of subjects in this data set are 1,010, we used 505 image pairs for CCA training in each fold.

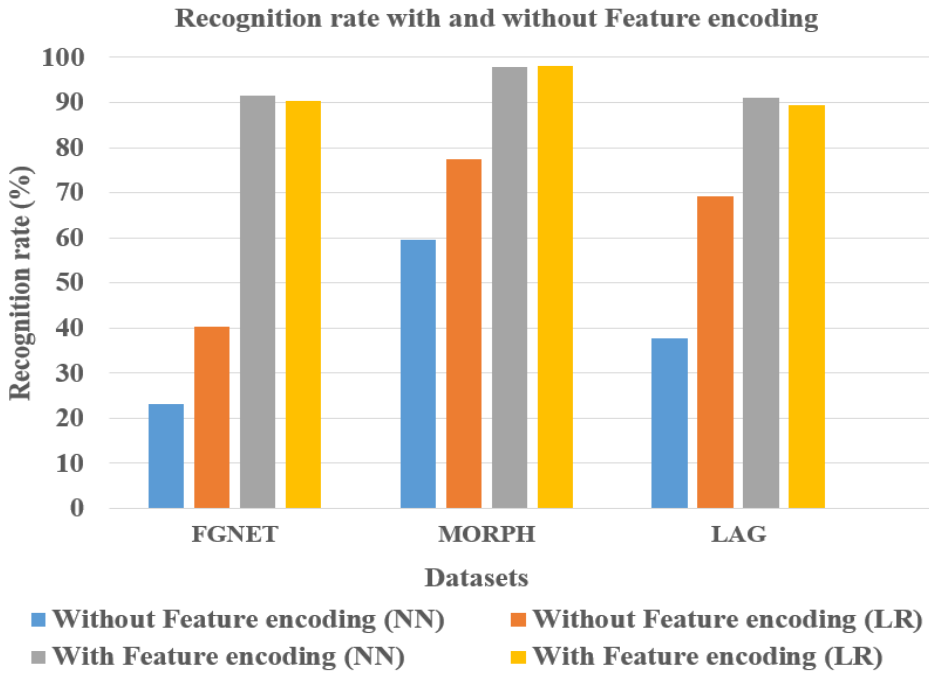


Fig. 4-11. Recognition rates with and without performing feature encoding for all the three data sets at the corresponding optimal feature dimensions.

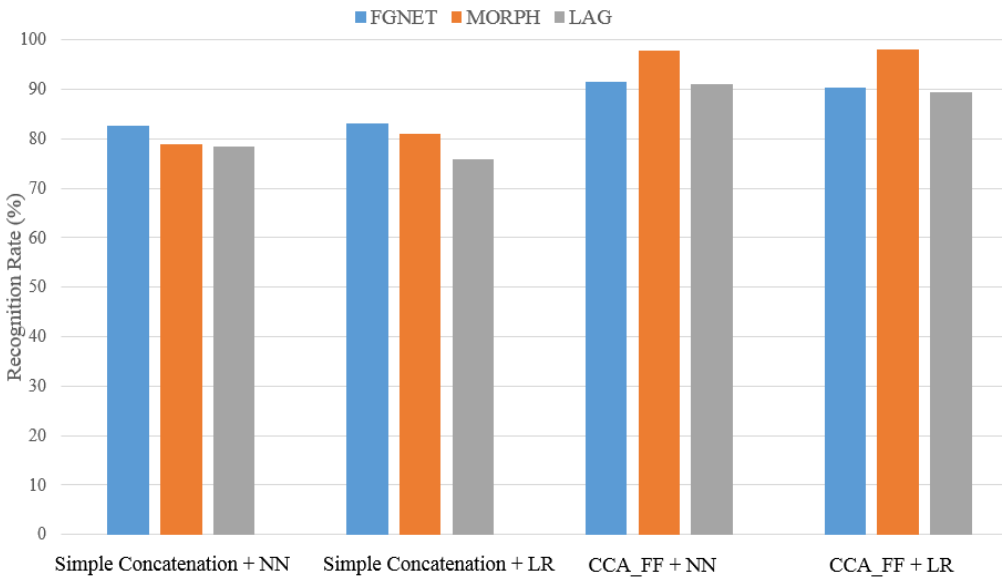


Fig. 4-12. Recognition rates with and without performing feature fusion using CCA for all the three datasets at the corresponding optimal feature dimensions.

4.6.5 Overall Benchmark Comparison

In our experiments, the performances of our proposed method are compared with state-of-the-art aging face recognition methods. For all the three datasets, the methods used for comparison are tuned to the same settings as used

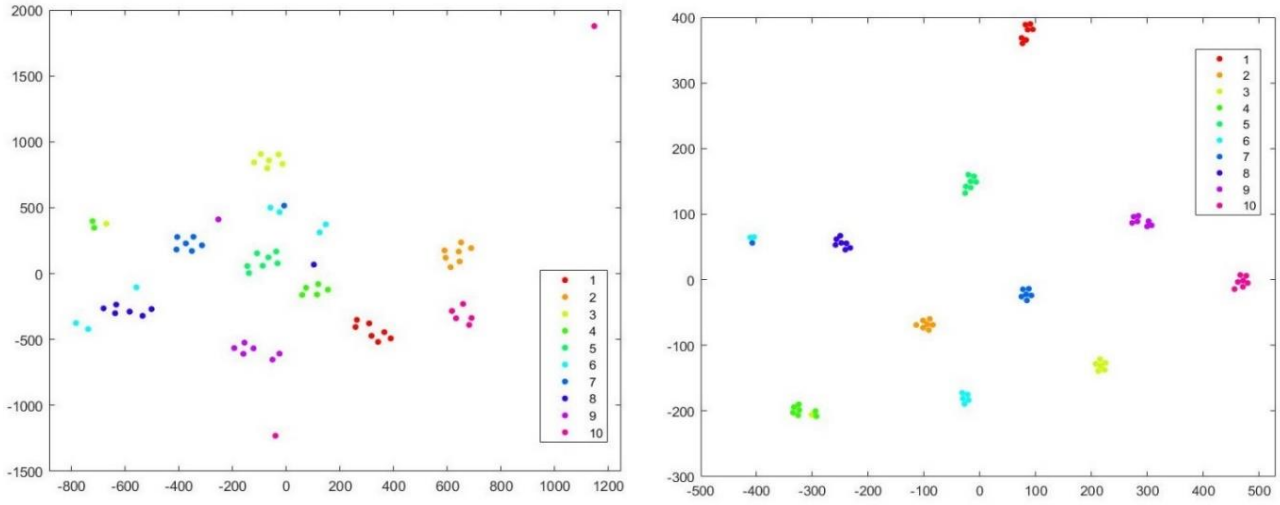
in their original literature, and the same protocol is used for the training and testing set. As discussed earlier, there is a large age gap between the images present in the gallery and the probe set, which are used for performance evaluation.

Due to large age variations, FGNET is the challenging data set for aging-face recognition. All the images were taken in the wild, with the age gap 0-45. Our proposed method provides superior results on this data set, which is 91.46% by using the deep features. Recognition results were recorded using the nearest neighbor (NN) classifier as well as linear-regression-based classifier. Furthermore, our proposed algorithm provides closed form solution for $\boldsymbol{W}$ using Equation (4.14), in which the matrix $\boldsymbol{P}$ is positive definite. On the other hand, analytical solution of $\boldsymbol{C}$ can be obtained as shown in (4.10). In comparison to the sparse-coding-based methods, the computational complexity of our proposed algorithm is relatively low, as it does not need to solve the l_1 minimization problem. Furthermore, the discriminant information of the training image pairs with a large age difference is further enhanced by exploiting the intra-person correlation, which is achieved by using CCA.

4.6.6 Better Reconstruction

According to the sparse-coding theory, each feature can be represented as a linear combination of multiple codewords for reconstruction. However, it is sensitive to feature variance, which may lead to selecting codewords far away from the input test feature in reconstruction. In this way, two similar features can select different codewords. Conversely, our method is based on locality information in searching nearest neighbors, so the selected local visual codewords can achieve better feature reconstruction. Our proposed method also leads to sparsity in the resultant encoded coefficients. Therefore, only the codebook entries near to the input (test) feature are selected for reconstruction. In other words, locality constraint ensures that the extracted features of the same person at different ages have similar codewords. As we discussed before, locality constraint captures the correlation among the features of the same identity by sharing the local bases of the codebook. However, this might not be possible for sparse coding, as non-zero sparse coefficients can be obtained for more than one class, especially in the presence of noise. The sparse coding approach can lead towards the sharing of dissimilar bases for the images of the same subject. Therefore, by incorporating both locality and sparsity terms in the objective function, performance will be affected along with the increase in computational complexity. Our proposed method learns the codebook using the Euclidean locality adaptor, which preserves the data structure for better feature representation, and recognition. Learning the codebook with the locality adaptor offers the following advantages: (1) non-smooth optimization can be avoided due to the use of a smooth objective function, and (2)

codebook size with locality adaptor has nothing to do with the data's dimension. For better understanding of our proposed feature encoding framework, we visualize the deep feature vectors, obtained before and after encoding, using t-distributed stochastic neighbor embedding (t-SNE) [168]. For visualization, we randomly selected 10 subjects with 7 images each from the LAG dataset. Fig. 4-13 show the deep features learned before and after applying our encoding framework. It can be seen that encoded features are well separated in the feature subspace. In other words, features of the same identity are represented by the same or similar codeword.



(a) (b)

Fig. 4-13. Visualization of the learned features before and after encoding using t-SNE. (a) before encoding, and (b) after encoding.

Our proposed feature-encoding-based method is different from LLC [122], due to the two main reasons. First, our formulation does not include the constraint $\|\boldsymbol{w}_k\|_2 \leq 1$, while LLC does. Due to the locality constraint$\|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2$, the size of the columns of $\boldsymbol{W}$ is not large. Removing the constraint $\|\boldsymbol{w}_k\|_2 \leq 1$ from the proposed formulation, we can obtain better optimized values, since a lesser number of constraints are included in our proposed objective function. In this way, the learned codebook $\boldsymbol{W}$ can efficiently capture the local structure of the data, which improves the recognition accuracy. Moreover, we can also obtain the closed form solutions for both the sparse coding phase and the codebook updating stage. Second, our proposed algorithm directly minimizes $\sum_{k=1}^{N} f(\boldsymbol{W}, \boldsymbol{f}_k)$, while LLC [122] minimizes $f(\boldsymbol{W}, \boldsymbol{f}_k)$, when the feature $\boldsymbol{f}_k$ is drawn from $\boldsymbol{F}$, and hence provides the approximated form of the objective function.

(a) (b) (c)

Fig. 4-14. Original images and noisy images obtained after adding Gaussian noise. (a) FGNET, (b) MORPH, and (3) LAG datasets.

4.6.7 Robustness to Noise Variations

In face recognition, a query face image may suffer from noise, due to various factors, such as environmental conditions, transmission errors, etc. These noise variations degrade the image quality, as well as the performance of face recognition systems. For comprehensive analysis, we evaluate the robustness of our proposed method, using a testing set contaminated by noise. Previously proposed methods [101, 102, 109] first decompose a facial image into identity, aging, and noise components, and then utilize the identity component for recognition. Xu et al. [109] proposed a coupled-auto encoder algorithm to first eliminate the noisy component from input images, and then perform recognition. It should be noted that these methods only consider the inherent noise in images, not the externally added noises, such as Gaussian noise, salt & pepper, etc. In our experiments, we add Gaussian noise into the probe face images, and then compute the recognition accuracy. The added Gaussian noise effects 30% of the pixels in a probe image. Our experimental results show the robustness of our proposed method in the presence of noise variations. Only 3-4% drop in recognition performance is observed for the MORPH and LAG data sets. However, a 20-25% decline in recognition rate is observed for the FGNET data set. The reason for this is that some images in the FGNET data set are of very poor quality, especially those childhood images. On the other hand, face images in the MORPH and LAG data sets have better quality, as compared to the FGNET data set. This problem can be solved by using some denoising filters, prior to feature extraction. Images obtained, after adding external Gaussian noise, are shown in Fig. 4-14. Fig. 4-15 shows the recognition results of our proposed method, with and without noise variations, using the nearest neighbor and the linear-regression classifiers [43].

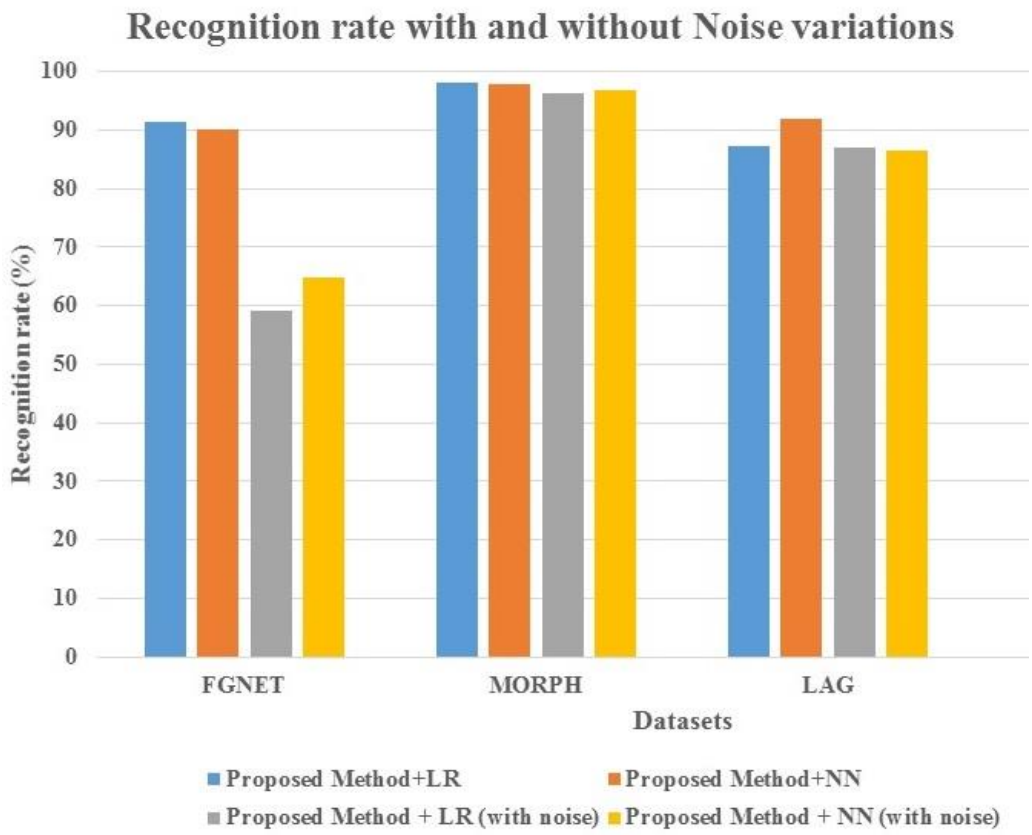


Fig. 4-15. Recognition rates of our proposed method on all the three data sets, with and without noise variations.

4.6.8 Computational Complexity Analysis

Our proposed method consists of three major parts, which are: (1) feature fusion using CCA, (2) codebook learning, and (3) feature encoding. In this section, we will evaluate the computational complexities of all the three stages of learning. As discussed before, the computational complexity of the feature encoding stage depends on the number of nearest neighbors selected for encoding. Results are reported with the optimal feature dimension and the number of nearest neighbors for all the three datasets. Table 4-4 shows the computation time in seconds for both feature fusion and the codebook-learning process. The computational complexity for the FGNET dataset is much lower than the MORPH and LAG dataset, as the size of the training set is small. The training set of the MORPH and LAG dataset consist of 20,000 and 4,000 images, respectively. For the MORPH dataset, the results are reported with a feature dimension of 40, and 150-NN. For the LAG dataset, the results are reported with the feature dimension of 40, and 100-NN. Table 4-5 shows the run time required to classify the whole testing set, as well as a single testing image. It can be observed that our proposed method can recognize a single testing image in less than one second, which shows that it is computationally efficient.

Table 4-4. Run time in seconds for the two stages of learning (Training).

Dataset	Feature Fusion using CCA	Codebook Learning
FGNET	0.011	0.8856
MORPH	0.264	54.8
LAG	0.12	16.10

Table 4-5. Run time in seconds for classifying one test image for all the three datasets (Testing).

Dataset	Feature encoding (single image)	Feature encoding (Whole dataset)
FGNET	0.0029	0.12
MORPH	0.0096	9.6
LAG	0.0052	0.26

4.6.9 Comparison with Local Feature Descriptors

In this section, the performance of our proposed feature-encoding framework is evaluated using two local feature descriptors, namely densely sampled scale invariant feature transform (DSIFT) [39] and local binary pattern difference (LBPD) feature [153]. Firstly, face regions are detected in an image using the Viola-Jones face detector [17], and are then resized to 150×200 pixels. To extract local features, an image is first divided into non-overlapping patches, and the selected feature descriptors are used to extract information from each of the patches. The extracted features are then concatenated to create a high-dimensional feature vector. In our proposed method, we perform dense sampling of the SIFT feature descriptors from the whole face image, which is equivalent to placing a regular grid on a face region as shown in Fig. 4-16.

Fig. 4-16. Placement of regular grid on a face image using DSIFT.

DSIFT is a robust local descriptor, which computes the local gradient information about image pixels. The neighborhood of each pixel $p = (x, y) \in I$ is divided into 4×4 patches, and each bin is represented by an 8-bin orientation histogram. This results in a $4 \times 4 \times 8 = 128$-dimensional vector for describing the pixel p. This operation

is performed for each pixel. Before extracting the dense SIFT features, we first perform image smoothing by convolving each face image with a Gaussian kernel of variance 0.25.

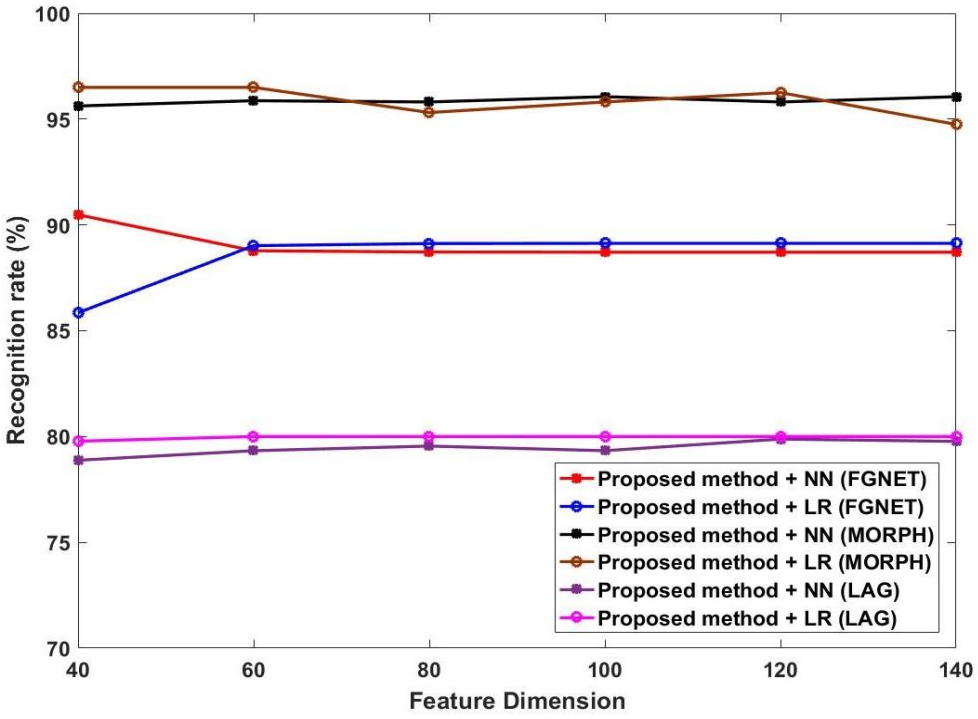


Fig. 4-17. Recognition rates with different feature dimensions using local feature descriptors (DSIFT+LBPD), and 150 nearest neighbours.

Recently, a numerical variant of LBP [153], known as local binary pattern difference (LBPD) was proposed, which offers several advantages over LBP. To extract the LBPD feature, the difference is computed between the LBP codes and its corresponding mean of a given local region. After extracting these two local features, we perform feature fusion to form a final feature vector, which is of very high dimensionality. To reduce the feature dimension, we utilize PCA, and then perform feature encoding using our proposed framework. Recognition rates with different feature dimensions for all the three data sets are shown in Fig. 4-17. It is observed that our method can also obtain a superior performance with local feature descriptors but there are some drawbacks. First, DSIFT features are computationally expensive, as it includes the gradient computation at each pixel. Secondly, handcrafted features are sensitive to noise, and their performance heavily depends on properly pre-processed face images.

4.6.10 Feature Selection and Fusion

The choice of features is always critical for face recognition systems. In our proposed method, we utilized DSIFT and LBPD features to extract useful information from face images. There are various reasons for that, which will be

described in this section. SIFT descriptor first detect the keypoints on face images, and then extract gradient information. For AIFR, this is not enough as human faces subject to large intra-personal variations, with age progression. Better recognition performance can be achieved by extracting information over dense grids instead of a few sparse keypoints. In this way, we can extract the information about the distribution of edge directions in the entire face region, which has been proved to be age-invariant discriminant information in [14]. Furthermore, using a single feature descriptor is not good enough to tackle the complex aging face recognition problem. Therefore, we fuse DSIFT with the LBPD feature to obtain a more discriminative and richer image representation. According to [178], LBP provides better recognition performance as compared to other local feature descriptors in terms of age progression. In the existing approaches of feature fusion, the correlation between the fused features is usually neglected, which results in a loss of useful information. Actually, LBP is produced by non-numerical responses, because the LBP codes are symbols or discrete patterns, so the Euclidean distance between two non-numerical features cannot be computed. Therefore, LBP codes cannot be combined or fused directly with other features, to form a new feature vector. In our experiments, we employ LBPD [153] feature (a numerical variant of LBP). It does not consider the intensity of pixels, because it utilizes the sign of comparisons between the neighboring pixels. This makes LBPD invariant to lighting conditions. As discussed in Section 3.3.1.4, one of its attractive properties is rotation invariance, as the norm used in Equation (3.7) and (3.8) makes sure that the code does not depend on permutation of the bits. The major difference between LBPD and LBP is that LBPD reflects the diversity of the local co-occurrence, instead of representing it directly. Furthermore, LBPD consists of numerical responses, whereas LBP is a collection of discrete patterns. This numerical property of LBPD makes it attractive in terms of texture analysis. Inspired by these properties, we utilize LBPD as a feature descriptor in our proposed framework.

4.7 Conclusions

In this Chapter, we have proposed a robust deep-feature-encoding-based method for age-invariant face recognition. Our method extracts high-level deep features using a pre-trained Deep-CNN model (AlexNet), and then performs feature encoding by using a Euclidean locality adaptor. At the training stage, we project pairs of training image's features into a coherent feature subspace using CCA, and then perform feature fusion. These fused features are then used to learn an age-discriminative codebook. At the testing stage, the query and gallery image's features are encoded using a learned codebook. For classification, the least squared method is first utilized to compute the

coefficient vector, which defines the relationship between the encoded gallery and query image's features. Based on that computed coefficient vectors, residual values are calculated for each class. Finally, a probe image is classified based on the least residual value. It is found that the locality information not only preserves the local strructure of the data samples, but also introduces sparsity in the resultant encoded coefficients. By incorporating the locality information in the codebook-updating process, same or similar codewords can be obtained for images of the same subject taken at different ages. Furthermore, closed form solutions can be obtained for both the encoding phase and the codebook updating stage. The proposed framework makes the feature more discriminative and robust to aging, as well as noise variations, as shown in our experiments. Experiment results on three challenging face-aging datasets show that our method performs better than conventional, as well as deep learning-based, age-invariant face recognition methods, and obtains the highest recognition accuracy.

Chapter 5 Deep Low-Rank Feature Learning and Encoding for Cross-age Face Recognition

5.1 Introduction

In Chapter 4, we proposed to solve cross-age face recognition problem using locality-constrained feature encoding framework. Furthermore, we demonstrated the robustness of our proposed approach against noise variations in a testing set. However, possible pixel corruption (random noise) in the training set was not considered. As discussed in the previous chapter, SRC [7] shows high robustness against occlusion, noise, and disguise. However, it assumes that the training data is taken under controlled environment without any noisy component. Therefore, the performance of SRC degrades heavily when training as well as testing data is corrupted with random noise. In Section 2.9, we briefly discuss the applications of low-rank matrix approximation techniques in context of face recognition. The previously proposed low-rank methods have shown high robustness to various noise variations on both training and testing data samples. However, these methods do not preserve the local structural properties of the data samples, when recovering the identity information of the face images. In Chapter 2, we briefly discussed the superiority of deep learning-based methods [48-54] in computer vision research, especially for face recognition. However, their performance is not optimistic in solving cross-age face recognition problem. The reason is that extracted deep features ignores the correlation among the images of the same identity taken at different ages. Furthermore, it is argued in [190] that deep models give worst performance, when a given test image suffered with noisy components. Till now, the performance of deep-learning models on noisy cross-age images is not investigated, which makes it a hot research topic.

In order to overcome the abovementioned challenges, we propose a robust deep low-rank feature-learning and encoding model, which consists of two stages of learning. Our method first extracts deep features from face images using a pre-trained deep CNN architecture (VGG16) [191], and then decomposes the extracted features into a low-rank feature matrix, and a sparse error matrix using our proposed manifold-constrained low-rank approximation algorithm. The learned low-rank features are then used to learn an age-discriminative codebook using our proposed feature-encoding framework. In the testing stage, the gallery and query image's low-rank features are encoded using a learned codebook. Finally, the nearest neighbor classifier is utilized to do face recognition. The encoded features are proved to be age discriminative as well as insensitive to noise variations, and provide state-of-the-art performance. Furthermore,

we have investigated the discriminative power of the periocular region of a human face for recognizing aging face images using our proposed framework.

The major contributions of this chapter are as follows:

- A novel low-rank decomposition algorithm based on local structural information is proposed to tackle noise variations in face images. We called it the manifold-constrained low-rank model. By learning a clean low-rank feature matrix using the proposed algorithm, better feature representation is obtained.
- A new feature-encoding framework based on exponential locality information is proposed, which converts learned low-rank features into an *N*-dimensional codeword by learning an age-discriminative codebook for final image representation. Our proposed method provides closed-form solutions in both the codebook-updating and encoding stages.
- Experimental results demonstrate the superiority of our proposed method over the other cross-age face recognition methods, in terms of recognition rate. The proposed low-rank decomposition algorithm recovers identity information from corrupted face images by removing sparse errors, and provides superior performance.
- Furthermore, we are the first one to investigate the cross-age face recognition problem in the presence of noise variations in both training and testing set.

The rest of the chapter is structured as follows. Section 5.2 present our proposed manifold-constrained low-rank approximation algorithm. In Section 5.3, we present our proposed feature encoding framework based on exponential locality constraint, which makes the extracted deep features more robust to age variations. Section 5.4 introduce the process of deep-feature extraction using a pre-trained deep-CNN model (VGG16). In Section 5.5, we explain the process of subspace learning and feature fusion using Kernel CCA. In Section 5.6, we present the experimental results of our proposed method with and without noise variations. Finally, Section 5.7 concludes our chapter.

5.2 Manifold-constrained low-rank matrix recovery

In this section, we will present our proposed manifold-constrained low-rank approximation function. Our major objective is to recover the underlying identity information from corrupted training and testing samples for better feature representation. Furthermore, learning low-rank features by incorporating manifold information has been proven to be

very helpful in generating a highly discriminative codebook for feature encoding. As our method is highly inspired by the low-rank approximation theory, so we will first discuss its fundamental mathematical formulation.

5.2.1 Robust PCA and Low-Rank Matrix Decomposition

In a real-world scenario, the data available for training and testing may be subject to occlusion, disguise or noise. Using these kinds of images for training and testing can significantly degrade the performance of face recognition systems, due to overfitting. Low-rank matrix recovery is a technique that has the ability to recover the clean data matrix from the given corrupted data sample, which can be used for recognition. Candes et al. [192] argues that the low-rank information can be retrieved under various conditions by utilizing the minimization problem with the nuclear norm. As discussed in Section 3.4, it minimizes the following nuclear norm regularized objective function:

$$\min_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 \;\; s.t. \boldsymbol{D} = \boldsymbol{A} + \boldsymbol{E}\,, \tag{5.1}$$

where $\boldsymbol{A}$ is a recovered low-rank matrix, $\boldsymbol{E}$ is an associated sparse error matrix, and λ is a parameter that controls the influence of the error term $\boldsymbol{E}$. Robust PCA [135] is one of the earliest proposed methods that recovers the underlying identity information from the corrupted data samples. To solve (5.1), several optimization algorithms have been proposed, such as accelerated proximal gradient (APG) [193], semi-definite programming (SPG) [194], etc. However, these methods have a large computational complexity, which makes them infeasible. Lin et al. [167] proposed the Augmented Lagrange multiplier (ALM) method, which shows promising results in solving the nuclear norm optimization problems with low-computational complexity. In fact, ALM is five times faster than APG [193], and it also requires less numbers of partial singular value decompositions (SVDs). As compared to APG, the number of non-zero entries produced by ALM in sparse error matrix $\boldsymbol{E}$ are small and precise. Therefore, we utilize ALM to solve our proposed constrained optimization problem.

5.2.2 Proposed Formulation

It is argued in [139] that the recovered low-rank matrix has the ability to preserve the global structural information of the data samples. To further enhance the discriminability of the features for representing data samples, we incorporate a manifold-constraint in our proposed objective function. The manifold information defines data samples by their nearest neighbors in the feature space, which preserves the local structure of the data samples. The proposed objective function is defined as follows:

$$\min_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 + \beta\mathcal{P}(\boldsymbol{A}) \tag{5.2}$$

$s.t. \boldsymbol{D} = \boldsymbol{A} + \boldsymbol{E}$

$$\min_{\boldsymbol{A},\boldsymbol{E}} \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 + \beta Tr(\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T)\,. \tag{5.3}$$

The first two terms represent the recovered low-rank matrix $\boldsymbol{A}$, and sparse error matrix $\boldsymbol{E}$, respectively. The third term learns the manifold structure of the data samples by computing a projection matrix $\boldsymbol{P}$ using locally linear embedding [30]. The reasons behind the selection of LLE in our algorithm can be found in Section 2.6.2. The parameter λ is used to control the impact of the sparse error term $\boldsymbol{E}$, and β is a manifold regularizer. The augmented lagrangian function to equation (5.3) is written as follows:

$$L(\boldsymbol{A},\boldsymbol{E},\boldsymbol{Y},\mu) = \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 + \beta Tr(\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T) + <\boldsymbol{Y}, \boldsymbol{D} - \boldsymbol{A} - \boldsymbol{E}> + \frac{\mu}{2}(\|\boldsymbol{D} - \boldsymbol{A} - \boldsymbol{E}\|_F^2)$$

$$= \|\boldsymbol{A}\|_* + \lambda\|\boldsymbol{E}\|_1 + f(\boldsymbol{A},\boldsymbol{E},\boldsymbol{Y},\mu) - \frac{1}{2\mu}\|\boldsymbol{Y}\|_F^2), \tag{5.4}$$

where <> represents the inner product, $\boldsymbol{Y}$ is a Lagrange multiplier, μ is a penalty parameter, $\|.\|_F$ is the Frobenius norm, and

$$f(\boldsymbol{A},\boldsymbol{E},\boldsymbol{Y},\mu) = \beta Tr(\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T) + \frac{\mu}{2}\left(\left\|\boldsymbol{D} - \boldsymbol{A} - \boldsymbol{E} + \frac{\boldsymbol{Y}}{\mu}\right\|_F^2\right). \tag{5.5}$$

Equation (5.4) is a constrained-minimization problem, where each variable is updated, while keeping the other fixed. In our proposed formulation, we have two variables $\boldsymbol{A}$, and $\boldsymbol{E}$, which are updated as follows:

Updating $\boldsymbol{A}$**:**

$$\boldsymbol{A}^{\boldsymbol{k+1}} = \arg\min_{\boldsymbol{A}} \|\boldsymbol{A}\|_* + <\nabla_{\mathrm{A}} f\left(\boldsymbol{A}^k, \boldsymbol{E}^k, \boldsymbol{Y}^k, \mu^k\right), \boldsymbol{A} - \boldsymbol{A}^k> + \eta\frac{\mu^{\mathrm{k}}}{2}\left\|\boldsymbol{A} - \boldsymbol{A}^k\right\|_F^2$$

$$= \arg\min_{\boldsymbol{A}} \|\boldsymbol{A}\|_* + <2\beta\boldsymbol{A}^k\boldsymbol{P} - \mu^k\left(\boldsymbol{D} - \boldsymbol{A}^k - \boldsymbol{E}^k + \frac{\boldsymbol{Y}^k}{\mu^k}\right), \boldsymbol{A} - \boldsymbol{A}^k> + \eta\frac{\mu^k}{2}\left\|\boldsymbol{A} - \boldsymbol{A}^k\right\|_F^2$$

$$= \arg\min_{\boldsymbol{A}} \|\boldsymbol{A}\|_* + \eta\frac{\mu^k}{2}\left\|\boldsymbol{A} - \boldsymbol{A}^k + \left[\frac{2\beta\boldsymbol{A}^k\boldsymbol{P}}{\mu^k} - \left(\boldsymbol{D} - \boldsymbol{A}^k - \boldsymbol{E}^k + \frac{\boldsymbol{Y}^k}{\mu^k}\right)\right]/\eta\right\|_F^2$$

$$\boldsymbol{A}^{k+1} = \frac{1}{\eta\mu^k}\left[\boldsymbol{A}^k + \left[-\frac{2\beta\boldsymbol{A}^k\boldsymbol{P}}{\mu^{\mathrm{k}}} + \left(\boldsymbol{D} - \boldsymbol{A}^k - \boldsymbol{E}^k + \frac{\boldsymbol{Y}^k}{\mu^k}\right)\right]/\eta\right] \tag{5.6}$$

Updating $\boldsymbol{E}$:

$$\boldsymbol{E}^{k+1} = \arg\min_{\boldsymbol{E}} \|\boldsymbol{E}\|_1 + < \boldsymbol{Y}^k, \boldsymbol{D} - \boldsymbol{A}^{k+1} - \boldsymbol{E} > + \frac{\mu^{\mathrm{k}}}{2} \left\| \boldsymbol{D} - \boldsymbol{A}^{k+1} - \boldsymbol{E} \right\|_F^2$$

$$= \arg\min_{\boldsymbol{E}} \frac{\lambda}{\mu^k} \|\boldsymbol{E}\|_1 + \frac{1}{2} \left\| \boldsymbol{E} - \left(\boldsymbol{D} - \boldsymbol{A}^{k+1} + \frac{\boldsymbol{Y}^k}{\mu^k} \right) \right\|_F^2$$

$$\boldsymbol{E}^{\mathrm{k}+1} = \frac{\lambda}{\mu^k} \left[\boldsymbol{D} - \boldsymbol{A}^{k+1} + \frac{\boldsymbol{Y}^k}{\mu^k} \right] \tag{5.7}$$

Algorithm 5-1: Manifold-constrained low-rank approximation

Input: Deep Features $\boldsymbol{D}$, regularization term λ, parameter β, and manifold projection matrix $\boldsymbol{P}$

Output: $\boldsymbol{A}^k$, $\boldsymbol{E}^k$

1**:** Initialize $\boldsymbol{A}^0 = \boldsymbol{E}^0 = \boldsymbol{Y}^0 = 0$, $\mu^0 = 0.1$

2: **while** not converged **do**

3: Update $\boldsymbol{A}$ using equation (5.6)

4: Update $\boldsymbol{E}$ using equation (5.7)

5: Update $\boldsymbol{Y}^{k+1}$ (Lagrange multiplier) as

$$\boldsymbol{Y}^{k+1} = \boldsymbol{Y}^k + \mu^k (\boldsymbol{D} - \boldsymbol{A}^{k+1} - \boldsymbol{E}^{k+1}) \tag{5.8}$$

6: Update k as $k + 1$

7: **end while**

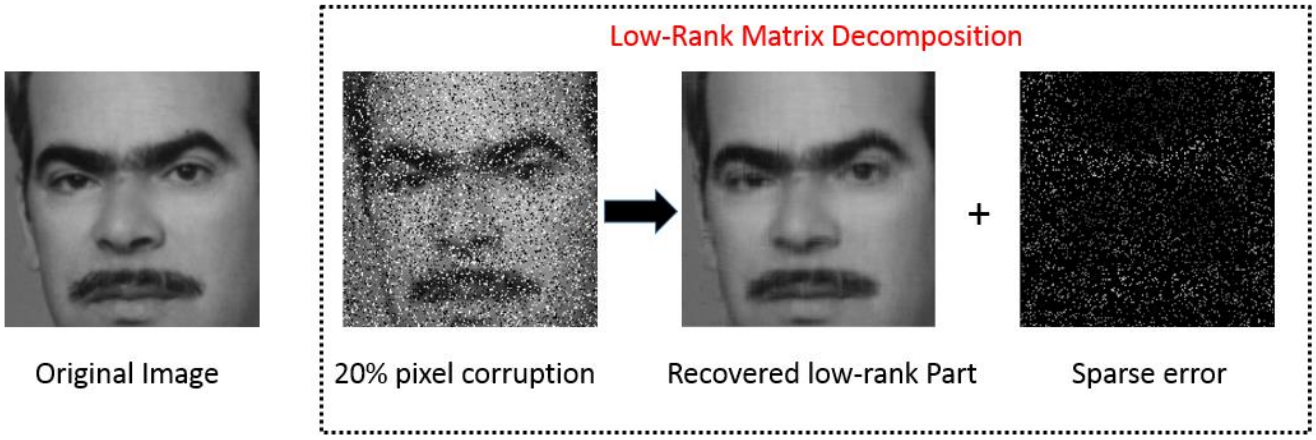


Fig. 5-1. Illustration of low-rank approximation algorithm.

In this model, the low rank matrix $\boldsymbol{A}$ contains the identity information, which is common among all the face images of a particular class, while the sparse error matrix $\boldsymbol{E}$ contains the information regarding facial variations, such as pose, lighting, expressions, age, etc. The learned low-rank matrix $\boldsymbol{A}$ exhibits two major properties; (1) correlation among the samples of the same class is enhanced due to manifold-constrained optimization; (2) geometrical structure of the samples in the learned low-rank feature space is preserved. Therefore, the learned representation is actually a combination of both global and local structural information, due to the use of low-rankness and manifold regularization, respectively. It is more suitable to learn the geometrical structure in the feature space to constraint the low-rank minimization problem. Therefore, we first learn the weight matrix $\boldsymbol{W}$ in the feature space using LLE, and then use it directly in our proposed objective function. The preserved geometrical information is depicted by the weight matrix $\boldsymbol{W}$. In this case, we minimize

$$\mathcal{P}(\boldsymbol{A}) = \sum_{i=1}^{n} \left\| \boldsymbol{a}_i - \sum_j w_{ij} \boldsymbol{a}_j \right\|_2^2 \tag{5.9}$$

$$= \|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{W}\|_2^2$$

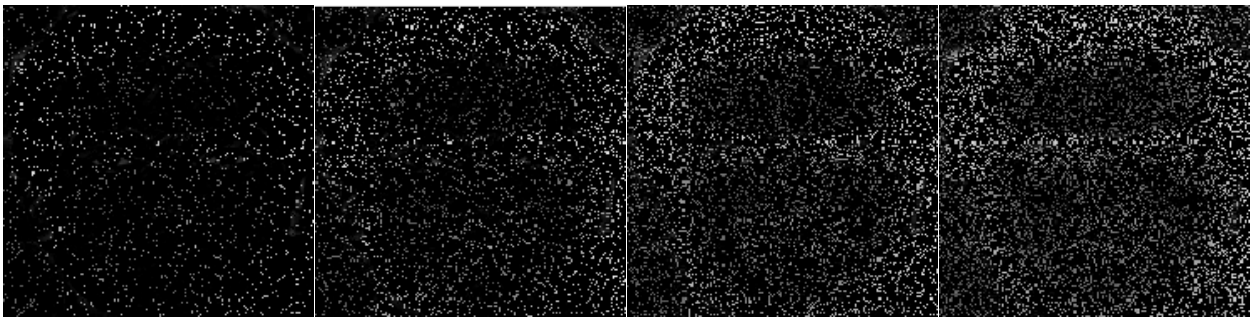
$$= \|(\boldsymbol{A} - \boldsymbol{A}\boldsymbol{W})^T\|_2^2$$

$$= Tr(\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{W})(\boldsymbol{I} - \boldsymbol{W})^T\boldsymbol{A}^T) = Tr(\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^T), \tag{5.10}$$

where $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{W})(\boldsymbol{I} - \boldsymbol{W})^T$ is the learned projection matrix. The approach used for learning the manifold subspace assumes that the data points and their corresponding neighbors lie on a linear patch, and its geometrical properties can be characterized by the linear coefficients. As compared to other manifold-learning algorithms, LLE is easy to implement because its optimization process does not contain local minima. Algorithm 5-1 summarizes our proposed manifold-constrained low-rank algorithm. Fig. 5-1 shows the decomposition of the corrupted face image into a recovered low-rank part and a corresponding sparse error matrix. Fig. 5-2 demonstrates the face images contaminated by different levels of noise variations, along with the recovered clean images and associated sparse errors.

(a)

(b)



(c)

Fig. 5-2. (a) Face images suffered from different levels of salt & pepper noise, (b) recovered low-rank images using the proposed algorithm, and (c) the corresponding sparse errors.

5.3 Low-rank Features Encoding based on locality information

We now explain our proposed feature-encoding framework based on the exponential locality constraint. Encoding the learned low-rank features provides better feature representation, which improves the recognition performance.

5.3.1 Locality information-based feature encoding framework

Consider that M low-rank feature vectors of dimension D are extracted, and they are represented as $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_M] \in R^{D \times M}$. The first step is to generate a codebook $\boldsymbol{Z}$ with N entries, i.e. $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_N\}$ using k-means clustering, which is then used to transform the extracted deep features into a N-dimensional

codeword for final feature representation. In other words, the extracted features are projected into a local coordinate system. According to the method proposed in [122], locality has been proven to be more important than sparsity. Feature encoding with locality constraint ensures that the resultant encoded coefficients are sparse. Our proposed locality-based objective function is defined as follows:

$$\min_{\boldsymbol{Z},\boldsymbol{C}} \|\boldsymbol{A}-\boldsymbol{Z}\boldsymbol{C}\|^2+\lambda\sum_{k=1}^{M}\|\boldsymbol{l}_k\otimes\boldsymbol{c}_k\|_2^2 \ , \tag{5.11}$$

$$\text{s.t.}\ \mathbf{1}^T\boldsymbol{c}_k=1,$$

where $\otimes$ represents the element-wise multiplication operator, $\boldsymbol{l}_k \in R^N$ represents the locality term that consists of an exponential locality adaptor, which provides freedom to each basis vector depending on its similarity to the given descriptor $\boldsymbol{a}_k$, and $\mathbf{1}$ represents the identity vector $[1,\dots,1]^T$. $\boldsymbol{C}=[\boldsymbol{c}_1,\boldsymbol{c}_2,\dots,\boldsymbol{c}_M]$ is the set of codes for $\boldsymbol{A}$. Each entry of the exponential locality term $\boldsymbol{l}_k$ can be defined as follows:

$$l_{kn}=\sqrt{\exp\left(\frac{\|\boldsymbol{a}_k-\boldsymbol{z}_n\|_2^2}{\sigma}\right)}, \tag{5.12}$$

where $\|\boldsymbol{a}_k-\boldsymbol{z}_n\|^2$ is the Euclidean distance between the input $\boldsymbol{a}_k$ and the n^{th} codebook entry $\boldsymbol{z}_n$. σ is a constant, which control the weight decay speed for the locality term. In our method, we choose $\sigma=0.5$. Later, we will present the justification behind the selection of this constant value. The locality term defined in (5.12) has an exponential growth with respect to $\frac{\|\boldsymbol{a}_k-\boldsymbol{z}_n\|^2}{\sigma}$. When the distance between $\boldsymbol{a}_k$ and $\boldsymbol{z}_n$ is large, the value of $\boldsymbol{l}_k$ becomes very large. As $\boldsymbol{l}_k$ is the weight of the sparse coefficient $\mathbf{c}_k$ in (5.11), large value of $\boldsymbol{l}_k$ leads to a small $\boldsymbol{c}_k$. Furthermore, the proposed locality constraint is quite different from the one proposed in [122], as it helps in deriving the closed-form solutions in the codebook-updating process. To solve (5.11), we utilize Lagrange multiplier, which is formulated as follows:

$$L(\boldsymbol{c}_k,\eta)=\|\boldsymbol{a}_k-\boldsymbol{Z}\boldsymbol{c}_k\|_2^2+\lambda\|\boldsymbol{l}_k\otimes\boldsymbol{c}_k\|_2^2+\eta(\mathbf{1}^T\boldsymbol{c}_k-1). \tag{5.13}$$

Let $\boldsymbol{B}=(\boldsymbol{a}_k\mathbf{1}^T-\boldsymbol{Z})^T(\boldsymbol{a}_k\mathbf{1}^T-\boldsymbol{Z})$, which is symmetrical. Equation (5.13) can be rewritten as:

$$L(\boldsymbol{c}_k, \eta) = \boldsymbol{c}_k{}^T \boldsymbol{B} \boldsymbol{c}_k + \lambda \boldsymbol{c}_k{}^T diag(\boldsymbol{l}_k)^2 \boldsymbol{c}_k + \eta(\mathbf{1}^T \boldsymbol{c}_k - 1), \tag{5.14}$$

where $\mathrm{diag}(\boldsymbol{l}_k)$ is a diagonal matrix. To determine the optimal solution of (5.14), we take its partial derivative and set it to zero, which gives the following equation:

$$\frac{\partial\big(L(\mathbf{c}_k, \eta)\big)}{\partial(\mathbf{c}_k)} = 2\boldsymbol{B}\boldsymbol{c}_k + 2\lambda diag(\boldsymbol{l}_k)^2 \boldsymbol{c}_k + \eta^T \mathbf{1} = 0 \tag{5.15}$$

Let $\boldsymbol{\Phi} = 2(\boldsymbol{B} + \lambda \mathrm{diag}(\boldsymbol{l}_k)^2)$, we have

$$\boldsymbol{\Phi}\boldsymbol{c}_k + \eta \mathbf{1} = 0, \tag{5.16}$$

Multiply (5.16) by $\mathbf{1}^T \boldsymbol{\Phi}^{-1}$, the following equation is obtained:

$$\mathbf{1}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\Phi} \boldsymbol{c}_k + \eta(\mathbf{1}^T \boldsymbol{\Phi}^{-1} \mathbf{1}) = 0. \tag{5.17}$$

According to the constraint $\mathbf{1}^T \boldsymbol{c}_k = 1$, $\mathbf{1}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\Phi} \boldsymbol{c}_k = 1$, we have:

$$1 + \eta(\mathbf{1}^T \boldsymbol{\Phi}^{-1} \mathbf{1}) = 0,$$

$$\eta = -(\mathbf{1}^T \boldsymbol{\Phi}^{-1} \mathbf{1})^{-1}. \tag{5.18}$$

Putting η into (5.16), we obtain the following equation:

$$\boldsymbol{\Phi}\boldsymbol{c}_k = (\mathbf{1}^T \boldsymbol{\Phi}^{-1} \mathbf{1})^{-1} \mathbf{1}$$

After some transformations, we have

$$\boldsymbol{c}_k = \frac{\boldsymbol{\Phi}^{-1}\mathbf{1}}{\mathbf{1}^T \boldsymbol{\Phi}^{-1} \mathbf{1}} = \frac{\frac{1}{2}\boldsymbol{\Phi}^{-1}\mathbf{1}}{\mathbf{1}^T \left(\frac{1}{2}\boldsymbol{\Phi}^{-1}\mathbf{1}\right)}. \tag{5.19}$$

In this way, the analytical solution of our proposed objective function can be obtained, which is formulated as follows:

$$\tilde{\boldsymbol{c}}_{\mathrm{k}} = \tfrac{1}{2}\boldsymbol{\Phi}^{-1}.\mathbf{1} = (\boldsymbol{B} + \lambda \mathrm{diag}(\boldsymbol{l}_{\mathrm{k}})^2)^{-1}\mathbf{1},$$

$$\boldsymbol{c}_k = \tilde{\boldsymbol{c}}_k/(\mathbf{1}^T\tilde{\boldsymbol{c}}_k). \tag{5.20}$$

The next step is to update the codebook $\boldsymbol{Z}$, which needs to solve the following equation:

$$\min_{\boldsymbol{Z}} \|\boldsymbol{A} - \boldsymbol{Z}\boldsymbol{C}\|^2 + \lambda \sum_{k=1}^{M} \|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|_2^2 \,. \tag{5.21}$$

Let the objective function, (5.21) be denoted as $\boldsymbol{A}(\boldsymbol{Z})$, which update the codebook generated using *k*-means clustering. Therefore, equation (5.21) finds the optimal codebook $\boldsymbol{Z}$ with $\boldsymbol{C}$ fixed. An analytical solution of (5.21) can be derived by taking the partial derivative of $\boldsymbol{A}(\boldsymbol{Z})$ with respect to the columns of $\boldsymbol{Z}$:

$$\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{z}_n} = \sum_{k=1}^{M} -2c_{kn}(\boldsymbol{a}_k - \boldsymbol{Z}\boldsymbol{c}_k) - 2\lambda \frac{\boldsymbol{l}_{kn}^2}{\sigma} c_{kn}^2(\boldsymbol{a}_k - \boldsymbol{z}_n) \tag{5.22}$$

In an equivalent form, it can be written as:

$$\left(\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{z}_n}\right)^T = \sum_{k=1}^{M} (-2c_{kn}\left(1 + \lambda \frac{\boldsymbol{l}_{kn}^2}{\sigma} c_{kn}^2\right)(\boldsymbol{a}_k)^T + 2(\lambda \frac{\boldsymbol{l}_{kn}^2}{\sigma} c_{kn}^2 \boldsymbol{z}_n{}^T + c_{kn} \sum_{j=1}^{N} c_{kj} \boldsymbol{z}_j{}^T)) \tag{5.23}$$

where $n \,\epsilon\, \{1,2,\dots,N\}$, and n is an index of the codebook entries. After setting the partial derivative of (5.21) to zero for $n = 1,2,3,\dots,N$, we obtain

$$g(\boldsymbol{Z}) = \boldsymbol{Z}\boldsymbol{R} - \boldsymbol{Q}^T, \tag{5.24}$$

where

$$\boldsymbol{R} = \sum_{k=1}^{M} \begin{pmatrix} (1+\lambda_{k1})c_{k1}^2 & c_{k1}c_{k2} & \cdots & c_{k1}c_{kN} \\ c_{k1}c_{k2} & (1+\lambda_{k2})c_{k2}^2 & \cdots & c_{k2}c_{kN} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1}c_{kN} & c_{k2}c_{kN} & \cdots & (1+\lambda_{kN})c_{kN}^2 \end{pmatrix}$$

$$\boldsymbol{Q} = \sum_{k=1}^{M} \begin{pmatrix} c_{k1}(1+\lambda_{k1}c_{k1})(\boldsymbol{a}_k)^T \\ c_{k2}(1+\lambda_{k2}c_{k2})(\boldsymbol{a}_k)^T \\ \vdots \\ c_{kN}(1+\lambda_{kN}c_{kN})(\boldsymbol{a}_k)^T \end{pmatrix}. \tag{5.25}$$

In this case, $g(\boldsymbol{Z})$ is a non-linear equation, so to search for the solution to $g(\boldsymbol{Z}) = 0$, we utilize Newton's method

$$\boldsymbol{Z}_{p+1} = \boldsymbol{Z}_p - g(\boldsymbol{Z}_p)(g'(\boldsymbol{Z}_p)^{-1}), \tag{5.26}$$

where $\boldsymbol{Z}_{p+1}$ is the learned codebook at the $p+1^{\text{st}}$ iteration. It can be verified that $g'(\boldsymbol{Z}_p) = \boldsymbol{R}_{\boldsymbol{Z}_p}$. In this way, we obtain

$$\boldsymbol{Z}_{p+1} = \boldsymbol{Z}_p - \left(\boldsymbol{Z}_p \boldsymbol{R}_{\boldsymbol{Z}_p} - \boldsymbol{Q}_{\boldsymbol{Z}_p}^T\right) \boldsymbol{R}_{\boldsymbol{Z}_p}^{-1} = \boldsymbol{Q}_{\boldsymbol{Z}_p}^T \boldsymbol{R}_{\boldsymbol{Z}_p}^{-1}. \tag{5.27}$$

Algorithm 5-2: Codebook updating scheme

Input: $\boldsymbol{Z}_{init} \in R^{D\times N}, \boldsymbol{A} \in R^{D\times M}, \sigma, \lambda$

Output: $\boldsymbol{Z}$

1: $\boldsymbol{Z} \leftarrow \boldsymbol{Z}_{init}$ (Initialize the codebook using *k*-means)

2: **for** $k = 1{:}\, M$ **do**

3: $\boldsymbol{l} \leftarrow 1 \times N$ (Locality constraint)

4: **for** $k = 1{:}\, N$ **do**

5: $\boldsymbol{l}_{kn} = \sqrt{\exp\left(\frac{\|\boldsymbol{a}_k - \boldsymbol{z}_n\|_2^2}{\sigma}\right)}$

6: end **for**

7: $\boldsymbol{l} \rightarrow normalize_{(0,1)}(\boldsymbol{l})$

8: Lagrange function to solve Equation (5.11)

$L(\mathbf{c}_k, \eta) = \|\boldsymbol{a}_k - \boldsymbol{Z}\boldsymbol{c}_k\|^2 + \lambda\|\boldsymbol{l}_k \otimes \boldsymbol{c}_k\|^2 + \eta(\mathbf{1}^T\boldsymbol{c}_k - 1)$

$\boldsymbol{c}_k = \tilde{\boldsymbol{c}}_k/(\mathbf{1}^T\tilde{\boldsymbol{c}}_k)$ Eq. (5.13) − (5.20) (Analytical solution)

9: Codebook Updating:

$$\left(\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{z}_n}\right)^T = \sum_{k=1}^{M}(-2c_{kn}\left(1 + \lambda\frac{\boldsymbol{l}_{kn}^2}{\sigma}c_{kn}^2\right)(\boldsymbol{a}_k)^T + 2(\lambda\frac{\boldsymbol{l}_{kn}^2}{\sigma}c_{kn}^2\boldsymbol{z}_n{}^T + c_{kn}\sum_{j=1}^{N}c_{kj}\boldsymbol{z}_j{}^T))$$

10: $\boldsymbol{Z}_{p+1} = \boldsymbol{Z}_p - \left(\boldsymbol{Z}_p \boldsymbol{R}_{\boldsymbol{Z}_p} - \boldsymbol{Q}_{\boldsymbol{Z}_p}^T\right) \boldsymbol{R}_{\boldsymbol{Z}_p}^{-1} = \boldsymbol{Q}_{\boldsymbol{Z}_p}^T \boldsymbol{R}_{\boldsymbol{Z}_p}^{-1}$

11: end **for**

12: Updated Codebook $\boldsymbol{Z}$

Therefore, the optimal solutions for encoding the parameters $\boldsymbol{C}$ and the codebook $\boldsymbol{Z}$ are obtained. This kind of iterative process is known as coordinate descent method. In this process, we optimize $\boldsymbol{C}(\boldsymbol{Z})$ based on the existing value of $\boldsymbol{Z}(\boldsymbol{C})$, alternatively. Finally, we can obtain the updated codebook by solving the non-linear system (5.24). In the testing stage, we perform fast encoding by searching for N_{knn} number of local bases from the codebook, which have minimum distances to the query image's feature vector. After that, the gallery and query image's features are encoded using the learned codebook $\boldsymbol{Z}$. The training and testing stages of our proposed method are summarized in Algorithms 5-2, and 5-3, respectively.

Algorithm 5-3: Feature encoding using a learned Codebook (testing stage)

1: **Input**: $\boldsymbol{Z}_n$ for $n \in \{1,2,\dots,N\}$ and $\boldsymbol{q}$ is a test image

2: for $k = 1 : N_{knn}$ **do**

where N_{knn} is number of nearest neighbors (number of local bases selected from the codebook)

3: First Euclidean locality adaptor $\boldsymbol{l}_k$ is computed

$$\boldsymbol{l}_{kn} = \sqrt{\exp\left(\frac{\|\boldsymbol{q}-\boldsymbol{z}_{kn}\|_2^2}{\sigma}\right)}$$

4: Solve the Equation (5.11) to obtain the analytical solution

$$\min_{\boldsymbol{c}_k} \|\boldsymbol{q} - \boldsymbol{Z}_{\boldsymbol{k}}\mathbf{c}_{\boldsymbol{k}}\|_2^2 + \boldsymbol{\lambda}\|\boldsymbol{l}_{\boldsymbol{k}} \otimes \boldsymbol{c}_k\|_2^2$$

$$\boldsymbol{\beta}^k = (\boldsymbol{B}_k + \lambda \text{diag}(\boldsymbol{l}_k)^2)^{-1}\mathbf{1}$$

$$\boldsymbol{c}_k = \boldsymbol{\beta}^k/(1^T\boldsymbol{\beta}^k)$$

where $\boldsymbol{B}_k = (\boldsymbol{q}\mathbf{1}^T - \boldsymbol{Z}_k)^T(\boldsymbol{q}\mathbf{1}^T - \boldsymbol{Z}_k)$

5.4 Deep Feature Extraction

Most of the face recognition algorithms extracts hand-crafted features from the face images before performing recognition. Although these features have achieved great success for some particular data, new-domain knowledge is required in order to learn some effective features for new data samples. Moreover, their performance is heavily affected in the absence of proper pre-processing operations. Recently, features

learned using deep-CNN models gains a lot of attention due to their ability of providing multiple levels of feature representation. In this regard, high-level features are assumed to provide valuable semantic information about the data samples. Furthermore, the learned deep-features are proved to be more invariant to intra-class variability. Therefore, deep-features are considered as a new class of effective feature-learning methods. In comparison to handcrafted features, deep-features reduce the amount of feature engineering, which makes them computationally efficient.

In our proposed method, we utilize a pre-trained deep CNN architecture, namely VGG16 [191], to extract the high-level features. Fig. 5-3 shows the VGG16 architecture. For VGG16, the input image must be of the size $224 \times 224 \times 3$. This deep architecture consists of five convolutional layers, having a receptive field of size 3×3. There are five max-pooling layers. Max pooling is performed over a window of size 2×2, with a stride of 2. Each convolutional layer is followed by a fully connected (FC) layer. In total, there are three FC layers in this deep architecture. The first two FC layers consist of 4096 channels, while the third contains 1000 channels. In our proposed method, we extract deep features from the second fully connected layer (fc7) for further processing. In deep learning-based methods, networks are trained for feature extraction and recognition from end to end. In our proposed approach, we use deep-learning-based CNN model to extract features, which are then converted into discriminative codewords for recognition. Figs. 5-4, and 5-5 demonstrates the features learned by the earlier and deeper convolutional layers of VGG16, from the entire face region, and the periocular region, respectively.

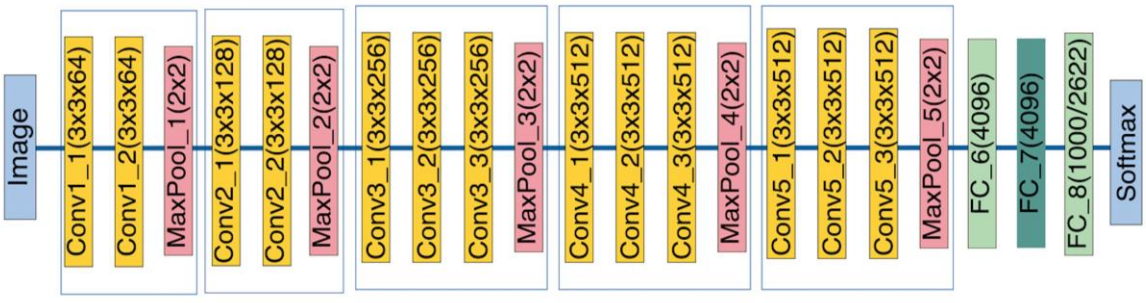


Fig. 5-3. VGG16 Architecture (Image adapted from [198]).

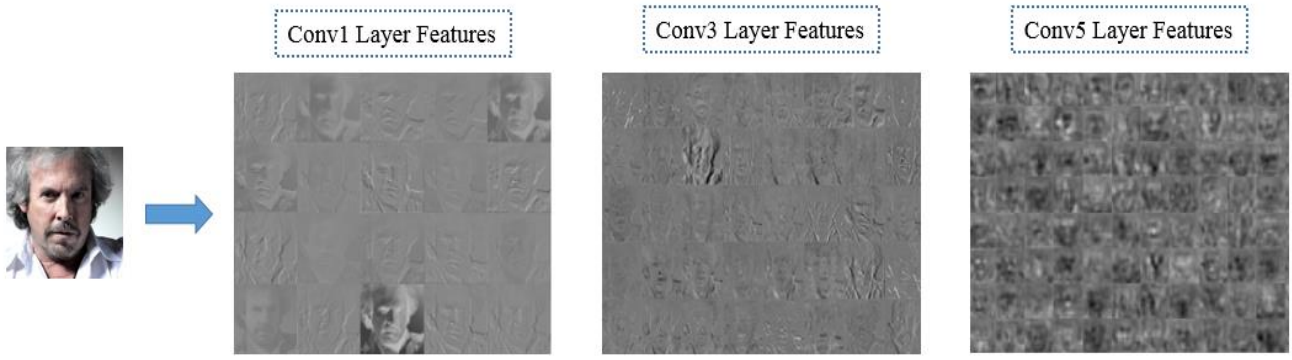


Fig. 5-4. Visualization of the learned deep features from different convolutional layers of VGG16.

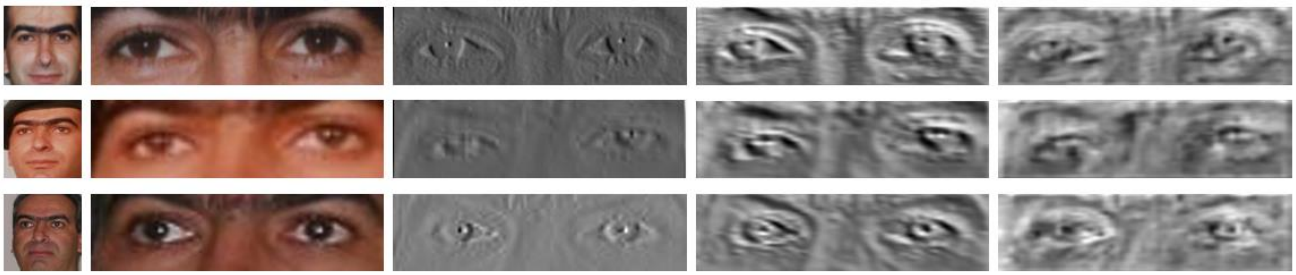


Fig. 5-5. (a) Original Face image, (b) Periocular region detected from face image, (c) Deep feature extracted from convolutional layer (conv1), (d) Deep feature extracted from convolutional layer (conv2), and (e) Deep feature extracted from convolutional layer (conv3).

5.5 Subspace learning and Feature Fusion using Kernel CCA

Features extracted from images of the same person at different ages are highly correlated, which can be exploited to enhance the recognition performance. As discussed before, the gallery and query image's features are encoded using a learned codebook, so the codebook must be age insensitive. The most common approach to determine the correlation between two sets of data samples is CCA, which seeks a linear transformation in such a way that the projected features in the transformed space have maximum correlation. The major drawback of CCA is that it cannot capture the nonlinear relations between the two samples. Therefore, in our algorithm, we utilize kernel canonical correlation analysis (KCCA) to learn the coherent feature subspace from pairs of training face images. In the training stage, we first divide the images of each subject into two parts. The first part contains images taken at younger ages, while the second part contain the images taken at elder ages. Therefore, one pair of face images (one at a younger age and the

other at an older age) is formed for each subject. We first project the extracted deep features of these image pairs of each subject into a coherent feature subspace, such that their correlation is maximized. These projected features are fused to form a final feature vector, which is then used to learn an age-discriminative codebook.

Given a pair of training image's features $\boldsymbol{F}_{age1}$ and $\boldsymbol{F}_{age2}$, we first perform mapping into the high-dimensional feature subspace by their corresponding mapping functions φ_{age1} and φ_{age2}, respectively. After mapping $\boldsymbol{F}_{age1}$ to φ_{age1}, and $\boldsymbol{F}_{age2}$ to φ_{age2}, we apply linear CCA, which moves it from primary to dual representation. The pair of directions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are learned to maximize the following criterion function.

$$\rho = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{\boldsymbol{\alpha}'\boldsymbol{K}_{age1}\boldsymbol{K}_{age2}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}'\boldsymbol{K}_{age1}{}^{2}\boldsymbol{\alpha}.\boldsymbol{\beta}'\boldsymbol{K}_{age2}{}^{2}\boldsymbol{\beta}}}, \tag{5.28}$$

where $\boldsymbol{K}_{age1} = \boldsymbol{F}_{age1}\,\boldsymbol{F}_{age1}{}^{T}$, and $\boldsymbol{K}_{age2} = \boldsymbol{F}_{age2}\,\boldsymbol{F}_{age2}{}^{T}$

Equation (5.28) can be maximized subject to two constraints $\boldsymbol{\alpha}'\boldsymbol{K}_{age1}{}^{2}\boldsymbol{\alpha} = 1$, and $\boldsymbol{\beta}'\boldsymbol{K}_{age2}{}^{2}\boldsymbol{\beta} = 1$, respectively. This can be solved using Lagrange multiplier method, which gives the corresponding equation:

$$L(\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}'\boldsymbol{K}_{age1}\boldsymbol{K}_{age2}\boldsymbol{\beta} - \frac{\lambda_\alpha}{2}\left(\boldsymbol{\alpha}'\boldsymbol{K}_{age1}{}^{2}\boldsymbol{\alpha} - 1\right) - \frac{\lambda_\beta}{2}\left(\boldsymbol{\beta}'\boldsymbol{K}_{age2}{}^{2}\boldsymbol{\beta} - 1\right) \tag{5.29}$$

Simplifying and solving Equation (5.29), we obtain

$$\boldsymbol{K}_{age1}\boldsymbol{K}_{age2}\boldsymbol{K}_{age2}{}^{-1}\boldsymbol{K}_{age1}\boldsymbol{\alpha} - \lambda^2\boldsymbol{K}_{age1}\boldsymbol{K}_{age1}\boldsymbol{\alpha} = 0$$

Hence,

$$\boldsymbol{K}_{age1}\boldsymbol{K}_{age1}\boldsymbol{\alpha} - \lambda^2\boldsymbol{K}_{age1}\boldsymbol{K}_{age1}\boldsymbol{\alpha} = 0$$

It can also be written as:

$$\boldsymbol{I}\boldsymbol{\alpha} = \lambda^2\boldsymbol{\alpha} \tag{5.30}$$

where $\boldsymbol{I}$ is an identity matrix. Finally, it becomes an Eigen-value problem of the form $\boldsymbol{Ax} = \lambda \boldsymbol{x}$. By doing this, we project the extracted low-rank features into the KCCA subspace for feature fusion. Figs. 5-6, and 5-7 shows the training and testing stages of our proposed feature-encoding framework.

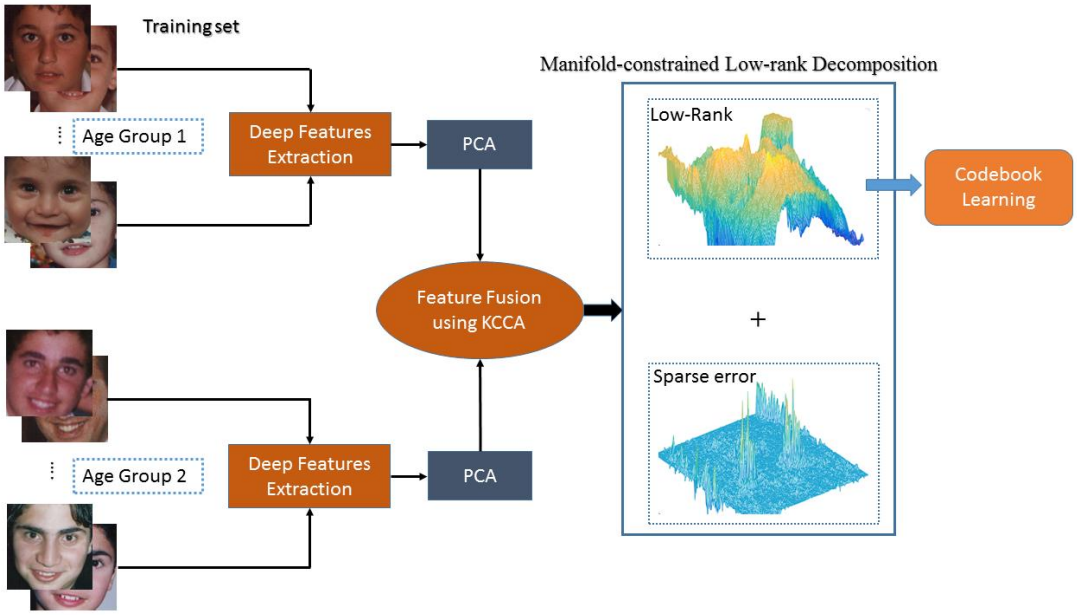


Fig. 5-6. Training stage of our proposed framework.

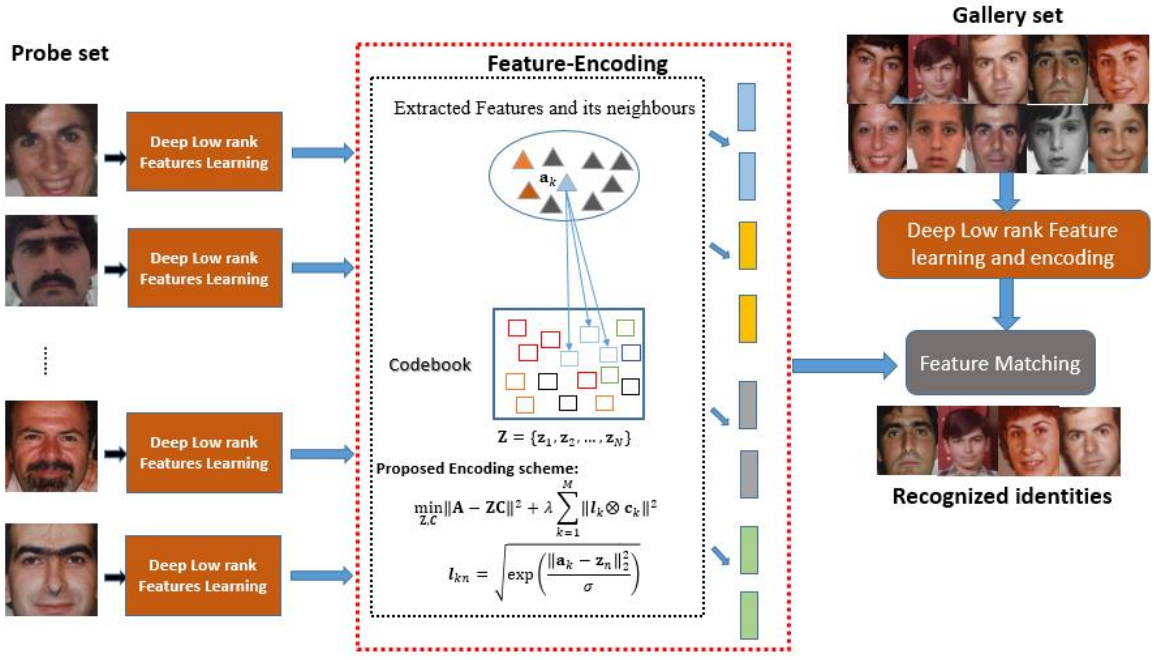
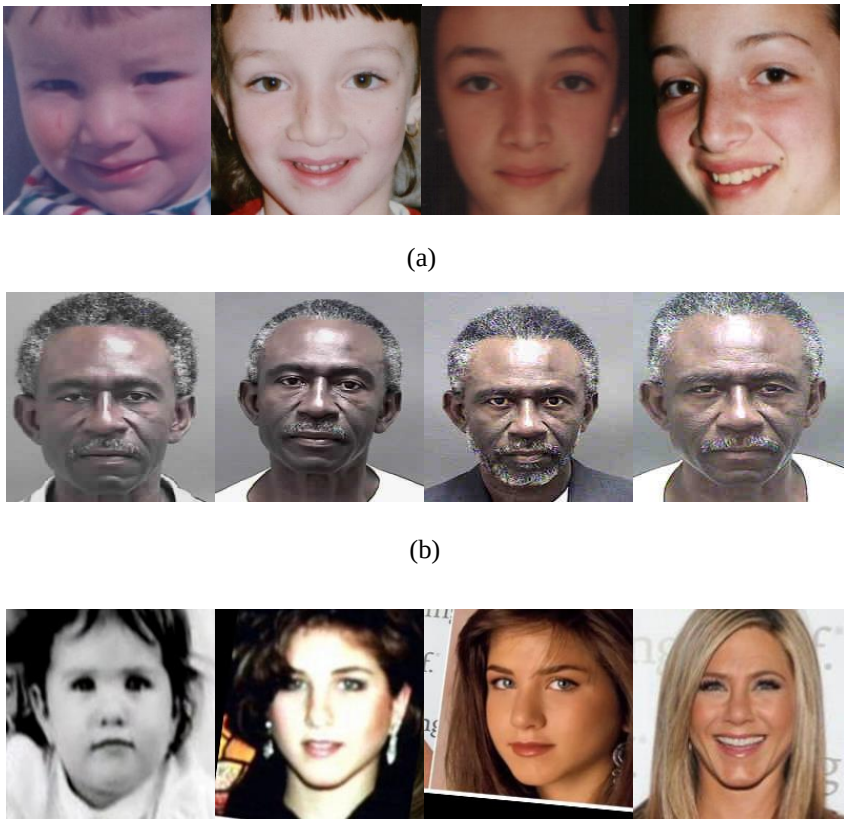


Fig. 5-7. Testing stage of our proposed framework.

5.6 Experimental Results and Analysis

The effectiveness of our proposed method is evaluated by conducting extensive set of experiments on three challenging face aging datasets. These include the FGNET [144], MORPH Album 2 [182], and Large-Age gap (LAG) [183] dataset. Among these datasets, FGNET and LAG contain images with very large age gap, which makes them challenging. For all three datasets, face regions are first detected using the Viola-Jones algorithm [17], and are then resized to $224 \times 224 \times 3$ pixels. Sample face images from the three datasets are shown in Fig. 5-8.



(a)

(b)

(c)

Fig. 5-8. Sample face images from the three face-aging datasets. (a) FGNET, (b) MOPRH, and (c) LAG dataset.

Our experiments are divided into two parts. Firstly, the whole face image is used for recognition. In the second part, the periocular region of a human face is first detected using Viola-Jones eye detector, which is then used for further processing. We compare the results of our proposed method with various state-of-the-art aging face recognition methods [14, 91, 101, 102, 103, 105, 106, 108, 109] in terms of recognition rate.

Furthermore, we also evaluate the performance of our proposed method with noise variations in both training and testing data. To the best of our knowledge, currently no literature exists for evaluating the cross-age face recognition problem in the presence of noise variations. Therefore, we perform some additional experiments on Multi-PIE dataset for comparative analysis.

5.6.1 Experiments on FGNET Dataset

FGNET is a small but challenging dataset, which consists of 1,002 face images from 82 identities with a large age difference of around 45 yrs. The number of images available per subject is large, around 10-12 images per subject. We follow the same experimental protocol as used in other papers, and evaluate the performance using the LOOCV scheme. Therefore, subjects included in the probe set are not in the training set. Sample images from FGNET dataset are shown in Fig. 5-8 (a). The training images of each subject are divided into two sets for KCCA pairwise training. As the age gap is quite large in this dataset, so the first set contains the younger age images, while the other one contains older age images. Following the LOOCV scheme, our training set contains N-1 total pairs, where N is the total number of subjects in the dataset. Comparative recognition rates are recorded in Table 5-1. Our proposed method achieves the recognition rate of 95.36%, which is better than the other aging face recognition methods. On the other hand, the recognition rate obtained by using the periocular region is 87.56%. Fig. 5-9 shows the face images from the FGNET dataset, and the corresponding detected periocular regions.

Fig. 5-9. Original face images and the corresponding detected periocular regions.

Table 5-1. Comparative results in terms of the rank-1 recognition rate on the FGNET Dataset.

Algorithms	Rank-1 Recognition rates
Park et al. [91]	37.4%
Li et al. [14]	47.5%
HFA [102]	69.0%
MEFA [101]	76.2%
CNN-baseline	84.4%
LF-CNN [106]	88.1%
Xu et al. [109]	86.5%
Proposed Method (Periocular region)	**87.56%**
Proposed Method (Whole face region)	**95.36%**

Table 5-2. Comparative results in terms of the rank-1 recognition rates on the MORPH database.

Algorithms	Rank-1 Recognition rate
Park et al. [91]	79.80%
Li et al. [14]	83.90%
HFA [102]	91.14%
MEFA [101]	92.26%
CARC [103]	92.80%
HOG+LPS [105]	94.20%
LPS [105] +HFA [102]	94.87%
LF-CNN [106]	97.51%
AFJT-CNN [108]	97.85%
Proposed Method (Periocular region)	**98.07%**
Proposed Method (Whole face region)	**97.93%**

5.6.2 Experiments on the MORPH Dataset

The MORPH dataset (Album 2) consists of 78,000 face images from 20,000 identities. The age gap among the images of the same subject is 5-6 yrs. Around 4 images per subject are available for each subject. In our experiments, we divide the dataset into independent training and a testing set. For training, the two

images with the largest age gap are selected from the first 10,000 subjects. The testing set is formed by using the remaining 10,000 subjects. Both the gallery set and the probe set are composed of one image per subject with a younger and an older age, respectively. This dataset is also suitable for addressing the small sample size problem, as only one image per subject is used in a gallery set. Most of the subjects in this dataset looks similar in terms of appearance, as they belong to same ethnicity. Sample images from MORPH dataset are shown in Fig. 5-8. (b). As training set consists of 10,000 subjects, so we utilize 10,000 image pairs for KCCA training. Our proposed method achieves the highest recognition rate of 97.93%, which is slightly better than the recently proposed deep-learning-based method in [108]. However, the recognition rate obtained using the periocular region is 98.07%. Table 5-2 shows the comparative analysis of our proposed method with respect to the other state-of-the-art aging face recognition methods.

5.6.3 Experiments on the LAG Dataset

LAG is a recently proposed face-aging dataset, which contain images with a very large gap, as shown in Fig. 5-8 (c). There are 3,828 face images from 1,010 identities. At least one child and one adult image are available for each identity, which makes it a very challenging dataset. For experiments, we follow the same experimental protocol as used in [183], and divide the dataset into two folds. Subjects are assigned to each fold alternately, so there is no overlapping between the training set and the testing set. The training set is formed by flipping the images in horizontal and vertical directions, and adding a certain amount of noise into each image. Therefore, we have 8 images for each subject in the training set. We conduct experiments for each fold independently, and then compute the average recognition rate. We compare the performance of our proposed method with two kinds of methods. The first kind of method employs DCNN [3] (trained on CASIA-Web face dataset [57]) for extracting deep features, and then use different metric learning techniques [181-184] for recognition. The second kind of method utilizes hand-crafted features (HDLBP [127]), with the abovementioned metric learning technique for recognition. Bianco et al. [183] fine-tuned DCNN [3], with the face images in the LAG dataset, to improve the performance. Therefore, we also include this method in our comparative analysis. Table 5-3 shows the comparative results, in terms of the

Rank-1 average recognition rate. With KCCA subspace learning and feature fusion, the training images of each subject are divided into two parts, with 4 images each. As the total number of subjects is 1,010, we utilized 505 image pairs from each fold for KCCA pairwise training.

Table 5-3. Comparative results, in terms of the rank-1 average recognition rates, on the LAG database.

Algorithms	Rank-1 recognition rate
DCNN [3]+SML [187]	72.43%
DCNN [3]+OSS [185]	66.42%
DCNN[3] + Cosine Similarity [184]	65.08%
DCNN [3] + Joint Bayesian [186]	66.33%
DCNN [3] + CARC [103]	74.82%
HDLBP [127]	71.53%
Bianco et al. [183]	84.95%
Proposed Method (Periocular region)	86.23%
Proposed Method (Deep features)	**92.56%**

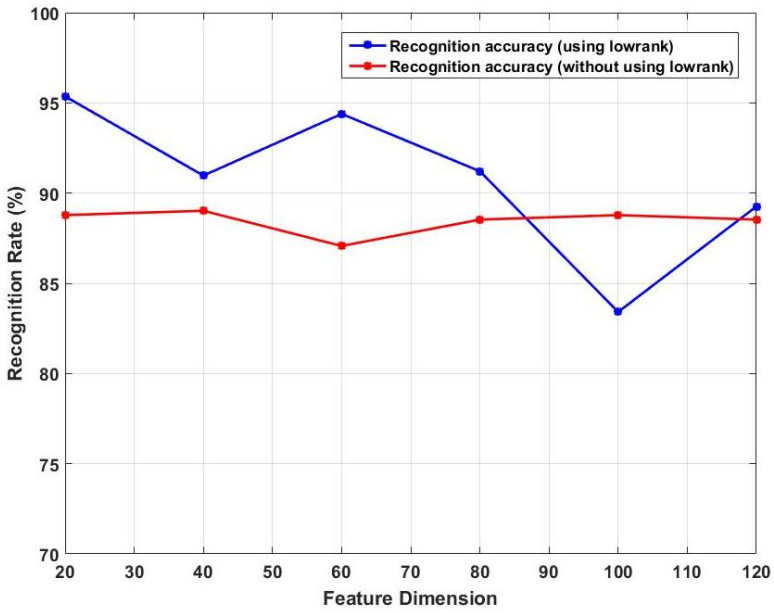


Fig. 5-10. The recognition rates under different feature dimensions, with and without using low-rank approximation on the FGNET dataset. (No noise)

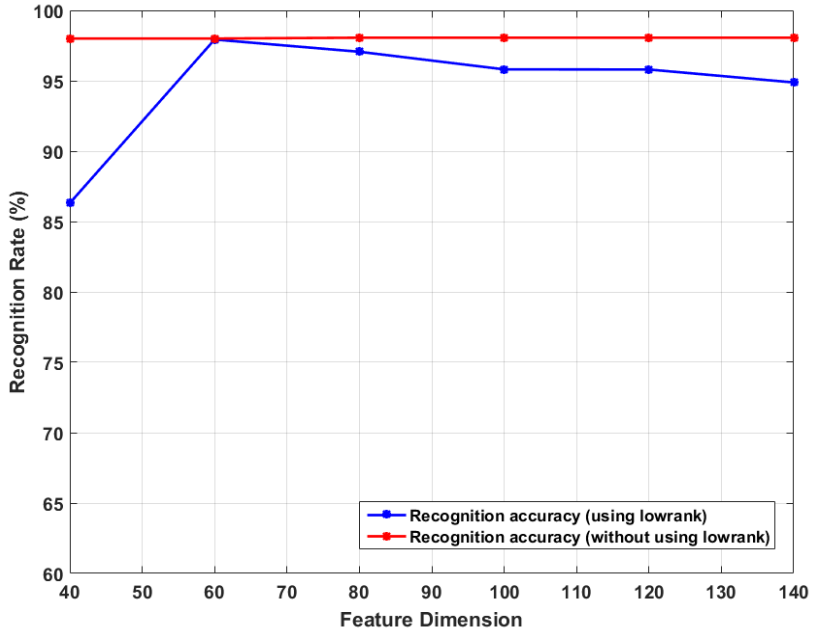
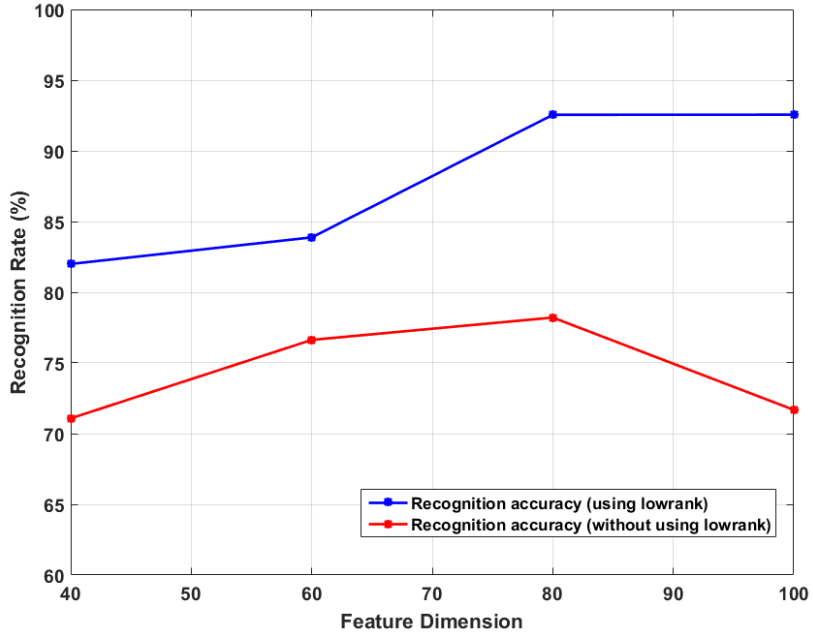


Fig. 5-11. The recognition rates under different feature dimensions, with and without using low-rank approximation, on the MORPH dataset. (No noise)



(c)

Fig. 5-12. The recognition rates under different feature dimensions, with and without using low-rank approximation, on the LAG dataset. (No noise)

5.6.4 Experiments on the CACD-VS Dataset

The CACD dataset is one of the largest available datasets for performing cross-age face recognition. It consists of 163,446 face images from 2,000 celebrities. All the images in this data set are collected from the Internet, with age labels. However, the whole data set consists of some incorrectly labelled samples. In addition to aging variations, there exist huge variations in terms of pose, expression, and illumination among the images of the same subject. The largest age-gap in this dataset is 10-12 yrs. For comprehensive analysis of our proposed method, we conducted experiments on a verification subset of CACD, named CACD-VS. The subset consists of 4,000 image pairs, including 2,000 positive pairs, and 2,000 negative pairs. All the images have been carefully annotated. Sample positive and negative pairs are shown in Fig. 5-13.

Fig. 5-13. Positive and negative pairs from the CACD-VS dataset, where first row represents the positive pairs and second row represents the negative pairs.

Following the same experimental protocol as used in [106], we evaluate the performance using the tenfold cross-validation rule. In this regard, we use cosine similarity as a metric learning technique for each pair and learn the optimal threshold value using nine training folds, for face verification. The leftover fold is used for testing. Experiments were repeated nine times, and average verification accuracy is reported in Fig. 5-14. The verification results are compared with different state-of-the-art cross-age face recognition methods. Our proposed method provides state-of-the-art performance, which is comparable to the deep learning-based method [106]. It is to be noted that the scale of the training data used by our method is smaller than the one used in deep learning-based methods.

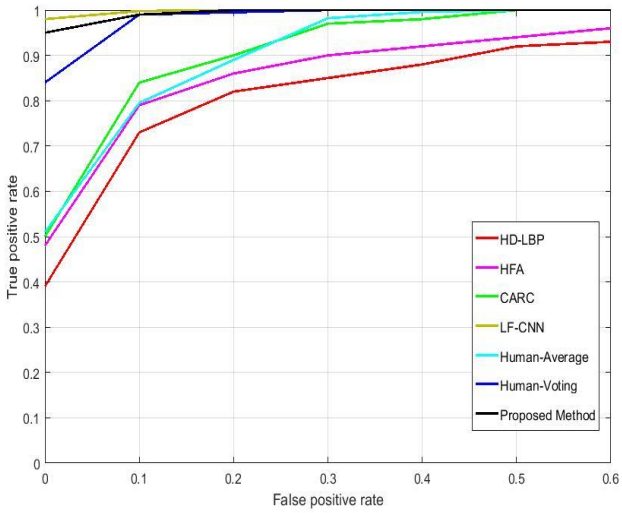


Fig. 5-14. Comparative analysis of ROC curves of different state-of-the-art methods.

Table 5-4. Statistics of the FGNET, MORPH (Album 2), and LAG face aging datasets.

Database	No of images	No of identities	Age range	Age gap	In wild
FGNET	1,002	82	0-69	0-45	Yes
MORPH (Album 2)	78,000	20,000	16-77	0-5	No
LAG	3828	1010	0-80	Large	Yes

5.6.5 Parameters settings for Low-rank Features Learning

Our proposed method includes two stages of learning. The first stage is the learning of the low-rank matrix $\boldsymbol{A}$ using Equation (5.6), which contains two important parameters λ, and β, respectively. Parameter λ is used to control the impact of the error term. In our experiments, we set $\lambda = 10^{-7}$, which implies that the amount of corruption in images could be constant. The second term β is used to control the impact of manifold regularization. As stated earlier, the projection matrix $\boldsymbol{P}$ is learned by using LLE [30], which preserves the neighborhood local structure of the data by representing the data samples in terms of their k nearest neighbors. By using a large training set, successful embedding can be obtained using LLE. However, the assumption of local linearity becomes invalid, when the size of the training set is small, and also a large number of nearest neighbors are used for the reconstruction of data samples. On the other hand,

using very small value of k may not be enough for successful embedding. The only way to find an optimal value of k is by means of iterative testing. In our experiments, we set k=60, which gives the optimum results. Furthermore, PCA is utilized to reduce the dimensionality of the extracted deep features. Therefore, we also assess the performance of our proposed method with different feature dimensions. Recognition rates for all the three datasets under different feature dimensions, without any noise, are shown in Figs. 5-10, 5-11, and 5-12, respectively. It can be observed that our method achieves state-of-the-art performance with lower feature dimensions, which is also an indication of lower computational complexity.

5.6.6 Parameters settings for Feature encoding

The second stage of our method is feature encoding based on locality information, which include the process of codebook learning. The total number of codebook entries is equal to the number of training samples. For the FGNET dataset, performance is evaluated using the leave-one-out scheme, so the codebook is initialized with N-1 entries, where N is the total number of subjects in the dataset. Due to the large number of training samples, the codebook size will be large. Therefore, the features of the gallery and query images encoded by using a large codebook will lead to a higher computational complexity. Inspired by LLC [122], we also utilize the approximate solution for fast encoding. Given a query image, we first compute the n entries in the codebook that have the minimum distances from the feature of the query image. In this chapter, we call these minimum-distance entries as the nearest neighbors, and the number n is one of the important parameters for this feature-encoding framework. These nearest neighbors are determined using kNN strategy proposed in [188]. In our experiments, we selected different numbers of nearest neighbors, ranging from 50 to 150, and computed the recognition accuracy. It is observed that, by increasing the number of neighbors, the recognition rate also increases. Our method can achieve superior performance by selecting only 100 nearest neighbors. We found that the recognition rate remains the same as we increase the number of neighbors from 100 to 150. Another important parameter is σ, which controls the locality error of the exponential locality adaptor, defined in Equation (5.12). If the value of σ is set such that

$\frac{\|\boldsymbol{a}_k - \boldsymbol{z}_n\|_2^2}{\sigma} < 1$, then we can obtain a lower error rate, as proved by Theorem 1 in [189]. Therefore, the value of σ should be as small as possible. In our experiments, we choose $\sigma = 0.5$, which gives the best recognition performance. Furthermore, the locality term $\boldsymbol{l}_{kn}^2$ defined in (5.12) has an exponential growth with respect to $\frac{\|\mathbf{a}_k - \mathbf{z}_n\|_2}{\sigma}$, which means that large value of $\boldsymbol{l}_{kn}$ is obtained if the distance between the input low-rank feature $\boldsymbol{a}_k$ and the codebook entry $\boldsymbol{z}_n$ is large. This implies the importance of data locality in the feature-encoding framework. Similarly, a large value of l_{kn} will cause the value of c_{kn} to be quite small, as l_{kn} is actually the weight of the encoding coefficient c_{kn}.

The value for the parameter λ in Equation (5.11) can be well explained by Equation (5.20). It should be noted that matrix $\boldsymbol{B}$ is semi-positive and symmetric. This implies that matrix $\boldsymbol{B} + \lambda \text{diag}(\boldsymbol{l}_k)^2$ is still conditioned, if $\boldsymbol{B}$ is singular or close to singular. In this way, the large distances will be penalized by $\lambda \text{diag}(\boldsymbol{l}_k)^2$, which analyze and capture the correlation between the two data points. In our experiments, we vary the value of λ from 10^{-9} to 10^{-3}, and found that by selecting $\lambda < 10^{-6}$, the resultant matrix becomes singular and provides inaccurate results. Therefore, we choose $\lambda = 10^{-3}$, which provides the best recognition performance.

5.6.7 Effectiveness of Low-Rank Feature Learning

Low-rank feature learning is proved to be an efficient technique to handle large amount of corruptions in the data samples. It aims to learn a low-rank dictionary by optimizing the dictionary atoms and removing the sparse noise errors from the data samples. Furthermore, it also reveals the global structural information of the data samples, which helps in reconstructing a given test sample using a discriminative low-rank dictionary. Learned low-rank features of the samples belong to the same class are highly correlated, which improves the classification performance. As discussed earlier, low-rank captures the global structural information of the data samples, so the correct identity of any data sample can be easily revealed, if subspaces are independent. In context of face recognition, low-rank matrix contains the identity information, while the sparse error term contains information regarding the facial variations. Therefore,

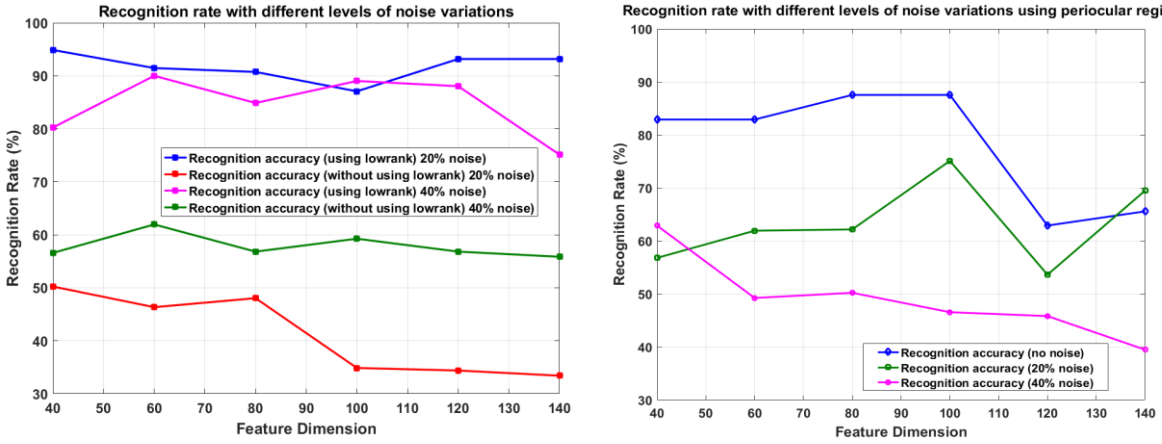
utilizing the low-rank part only is quite beneficial for recognition purposes. The discriminative power of the low-rank dictionary can further be enhanced by incorporating the local structural information into account. By utilizing the manifold information in our proposed method, data samples of the same identity lie close to each other in the learned subspace, which improves the recognition rate. Therefore, the combination of the global and local structural information provides better feature representation.

5.6.8 Evaluation with Noise Variations

Practical face recognition systems must be able to handle noise variations in images for reliable identification. As we discussed earlier, deep neural networks are highly sensitive to noise variations. Therefore, we propose a manifold-constrained low-rank decomposition algorithm to recover underlying identity information from corrupted face images for face recognition. The sparse-representation based classifier (SRC) [7] provides superior performance in recognizing face images suffering from occlusion, disguise, and noise. However, SRC does not consider any possible contamination in the training set. Therefore, its performance will degrade heavily when training face images are corrupted. In our proposed method, we consider both the training and testing data to be contaminated with the salt & pepper noise, with different levels of pixels corruption. After decomposing the deep features of a noisy image into a low-rank feature and a sparse error matrix, we only utilize the low-rank component for recognition, while discarding the sparse error matrix. In our experiments, we found that even if no random noise is added to an image, there still exist some inherent random noise, which is generated due to the camera-acquisition system. Previously proposed methods have been proved to be robust against such kind of noise, and provide superior performance. However, their performance under large externally added noise is not investigated yet.

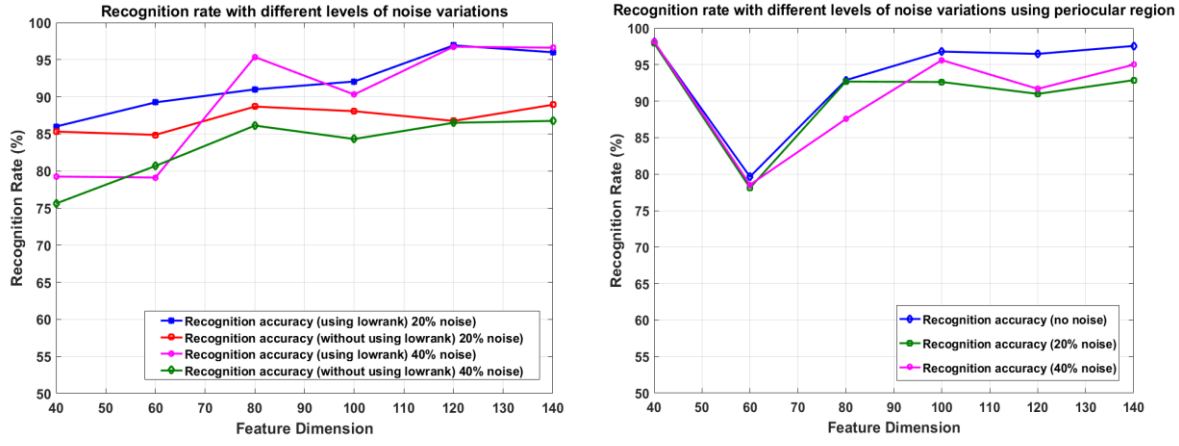
In addition to recovering the identity information, low-rank approximation is also capable of alleviating illumination variations, as shown in Fig. 5-2 (b). In our experiments, we randomly add 20 to 40% “salt & pepper” noise to the training and testing images. We evaluate the performance, with and without using our proposed low-rank approximation technique, under different feature dimensions on the FGNET, MORPH,

and LAG datasets, as shown in Figs. 5-15(a), 5-16(a), and 5-17(a), respectively. It can be observed that our proposed low-rank algorithm shows high robustness to different levels of pixel corruption. This is because of the utilization of the recovered clean feature matrix for feature encoding. Face images of the same identity are linearly correlated, so the learned codebook used to represent images from one class should be of low rank. Furthermore, the introduction of manifold regularization not only preserves the local structural information of the data samples, but also provides a compact and clean dictionary. Therefore, it is able to reconstruct clean images from noisy observations, even in the presence of corrupted training data.



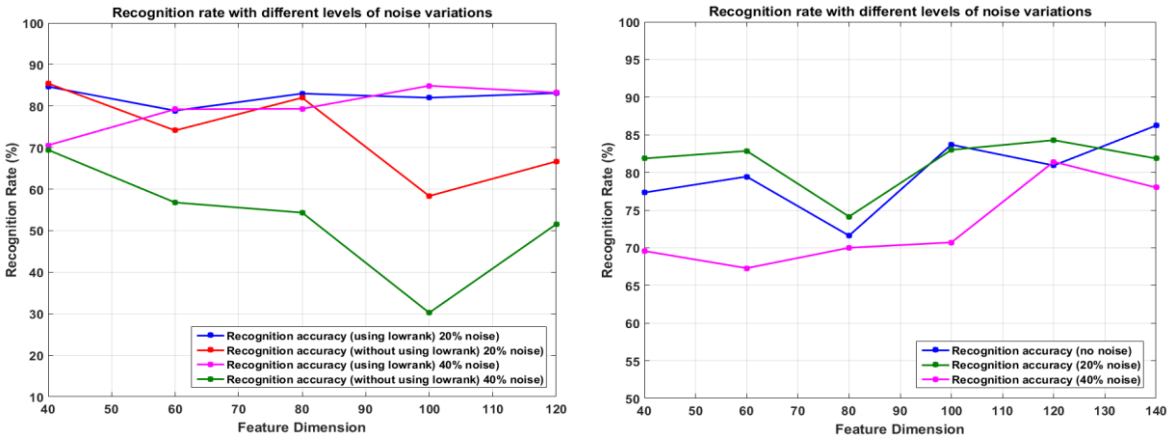
(a) (b)

Fig. 5-15. Recognition rates under different feature dimensions, with different levels of noise variations, on the FGNET dataset. (a) whole face region, and (b) the periocular region.



(a) (b)

Fig. 5-16. Recognition rates under different feature dimensions, with different levels of noise variations, on the MORPH dataset. (a) whole face region, and (b) the periocular region.



(a) (b)

Fig. 5-17 . Recognition rates under different feature dimensions, with different levels of noise variations, on the LAG dataset. (a) whole face region, and (b) the periocular region.

Considering the whole face image, our proposed method achieves superior results on all three face-aging datasets. For the FGNET dataset, our method achieves the recognition accuracy of 93.17%, when 20% of the pixels are corrupted by noise. Only a 3% drop in recognition rate is observed when we increase the number of contaminated pixels from 20% to 40%. For the MORPH dataset, only a 1% decline in recognition rate is observed with 20, and 40% of pixels corrupted using salt & pepper noise. In the second part of our experiments, we investigated the use of the periocular region for face recognition on all the three datasets, with different noise levels, and reported the results in Figs. 5-15(b), 5-16(b), and 5-17(b), respectively. Although the performance of using the periocular region is not the same as using the whole face image, it still shows high robustness against noise variations. For the FGNET dataset, only a 10% decline in recognition rate is reported when 20% of the pixels are corrupted, while 20% decline is reported for 40% corrupted pixels. The reason for this is that most of the childhood images in the FGNET dataset were actually scanned from the original photographs, captured under large pose, illumination, and expression variations. However, face images in the MORPH dataset are captured under controlled conditions, so the periocular region shows high robustness to noise variations. For the LAG dataset, the results are slightly better than FGNET, due to the better image quality. However, under normal circumstances (no noise), using

the periocular region has proven to be highly discriminative and provides a recognition rate better than the other state-of-the-art cross-age face recognition methods.

5.6.9 Computation time

In this section, we will evaluate the computational efficiency of our proposed framework. The computation time depends on the two stages of learning, including manifold-constrained low-rank approximation, and feature encoding. For low-rank optimization, we measured the runtime for both a single image and the whole dataset. In the testing stage, our proposed method takes only 0.2 seconds to reconstruct the clean image from the noisy observation. The main computation in Equation (5.6) is updating the matrix $\boldsymbol{A}$, which involves the process of singular value decomposition (SVD) of an $M \times M$ matrix. For the feature-encoding stage, the computational complexity depends on the number of local bases (nearest neighbors) selected from the learned codebook for encoding. Table 5-5 shows the runtime required for the feature-encoding stage. Our proposed method takes only a few milliseconds to perform feature encoding.

Table 5-5. Computation time in seconds for encoding one single image and the whole data set.

Dataset	Feature Encoding (Single image)	Feature encoding (whole testing dataset)
FGNET	0.0029s	0.12s
MORPH	0.0096s	9.60s
LAG	0.0052s	0.26s
CACD-VS	0.0048s	3.84s

Fig. 5-18. Sample face images from Multi-PIE Dataset.

5.6.10 Comparative Analysis on Multi-PIE dataset

As discussed earlier, we are the first one to investigate the recognition performance on cross-age face datasets in the presence of noise variations. For comparative analysis, we perform some additional

experiments on Multi-PIE [169] dataset, with both training and testing images to be corrupted with different levels of salt & pepper noise. Earlier in this chapter, we discussed some limitations and drawbacks of sparse representation-based classifier (SRC), in solving the noisy and occluded face recognition problem, when training set is corrupted with noise. To overcome the limitations of SRC, Jiang et al. [195] proposed a method, known as sparse and dense hybrid representation (SDR). The method first splits the training set into a class-specific dictionary and a non-class-specific dictionary to obtain maximum amount of information, so it can deal with the corrupted pixels in the face images. Although, this method provides satisfactory performance in dealing with little amount of corruption, but its performance heavily degrades, when heavily corrupted face images are presented for recognition. Moreover, the computational complexity of SDR depends on the size of training dictionary. Large training data leads to a very high computational complexity, which is infeasible.

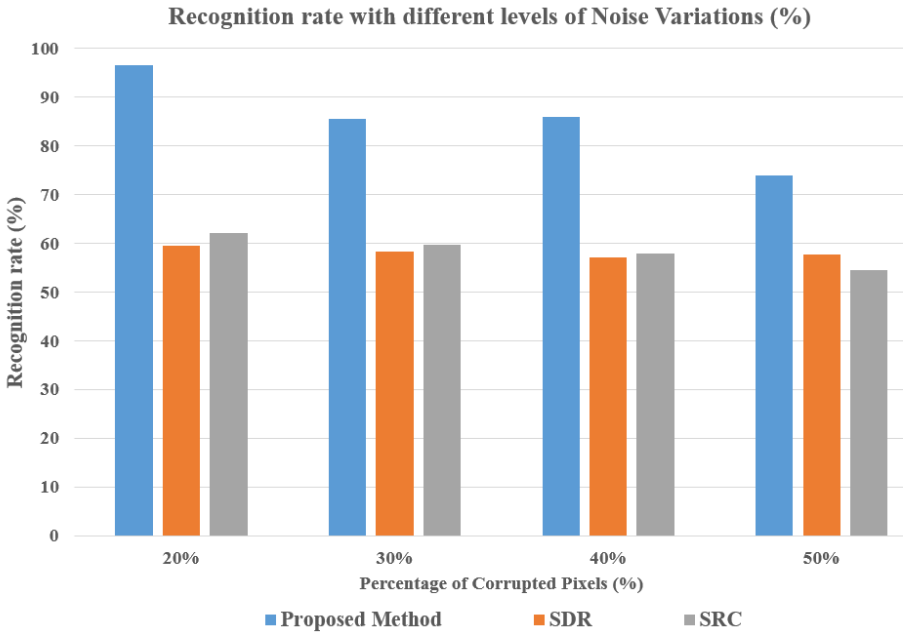


Fig. 5-19. Recognition rates with different levels of noise variations using deep features on Multi-PIE Dataset.

Multi-PIE consists of more than 750,000 images from 337 subjects. In our experiments, we randomly select 68 subjects with 102 images each. For training, we randomly select 34 subjects, while the remaining subjects are used for testing. Sample images from the Multi-PIE dataset with different facial variations are

shown in Fig. 5-18. In our experiments, gallery set is assumed to be free of external noise, while training set and probe face images are corrupted with 20, 30, 40, and 50% of salt & pepper noise, respectively. With 20% corrupted pixels, our method provides state-of-the-art performance by achieving the recognition rate of 96.60% with 60-D features. With 50% of the corrupted pixels, our method can still achieve the recognition rate of more than 70.00% with 60-D features. Comparative results are shown in Fig. 5-19. In our experiments on face-aging datasets, we divide the images of each subject into two parts with large age gap. As Multi-PIE is not a face-aging dataset, so we randomly divide the images of each subject into two parts, with 51 images each, and then use them for KCCA pairwise training.

5.6.11 Evaluation with Local feature descriptors

In this section, we evaluate the performance of our method using local feature descriptors, with different levels of corrupted pixels in the training and testing set. In our experiments, we add 20, 30, and 40% of salt & pepper noise in the training and testing images. To extract the features, we first detect the location of the face region in an image using the Viola-Jones face detector [17], and then resize the face region to 150×200 pixels. Each face image is first smoothed by using a Gaussian kernel with a variance of 0.25. As discussed in Chapter 4, extracting information over dense grids instead of a few sparse key points can provide the information regarding the distribution of edge directions in the entire face region. According to [14], this information is proved to be age-invariant. Regarding local feature descriptors, our experiments are divided into three parts. In the first part, we extract only Dense SIFT (DSIFT) features from the face images, using the same settings as used in Chapter 4. Extracted features are then passed to our proposed low-rank and feature-encoding algorithm to obtain better feature representation.

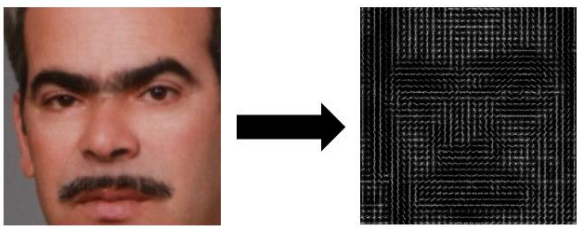


Fig. 5-20. Sample face images and its corresponding extracted HOG feature.

In the second part, we utilize two efficient local features, HOG [36], and LBPD [153]. To extract HOG feature, we first divide an image into small cells, and then compute the gradient information at each cell, respectively. In our experiments, we set the size of each cell to 8, while the number of orientations used in the histogram are set to 4. LBPD feature is extracted in a same way as in previous chapters. Finally, we concatenate the extracted HOG and LBPD feature to extract more discriminative face representation. The HOG feature extracted from a face image is shown in Fig. 5-20. In the last part, we fuse LBPD and DSIFT features, as explained in Chapter 4, and record the recognition rate using our proposed method. In our experiments, we found that feature-level fusion brings a significant improvement in recognition performance. The best performance is obtained by fusing LBPD with DSIFT and HOG features. For all the three combinations of local features, the highest recognition rate is obtained using 40-D features, and by searching for 150 nearest neighbors (local bases) from the codebook for feature-encoding. In comparison to local feature descriptors, deep-features performs better against different levels of pixel corruptions, as shown in Figs. 5-15, 5-16, and 5-17 (a), respectively. Fig. 5-21 shows the recognition rates of our proposed method of all the three datasets using only DSIFT feature. Similarly, recognition rates using multiple local features are shown in Figs. 5-22, and 5-23, respectively. Tables 5-6, 5-7, and 5-8 shows the recognition results under different feature dimensions using DSIFT and LBPD feature with different levels of noise variations.

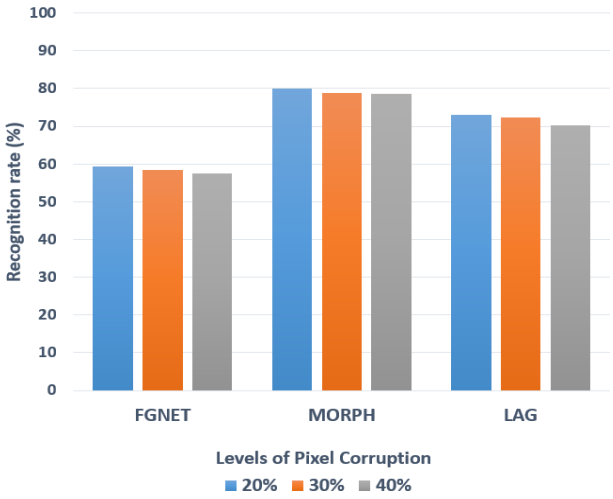


Fig. 5-21. Recognition rates of our proposed method using DSIFT feature, with optimal feature dimensions.

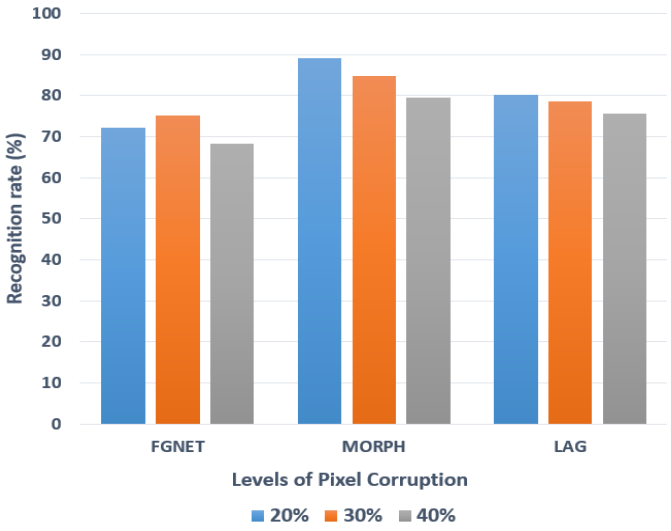


Fig. 5-22. Highest recognition rates of our proposed method using local feature descriptors (HOG+LBPD) with optimal feature dimensions.

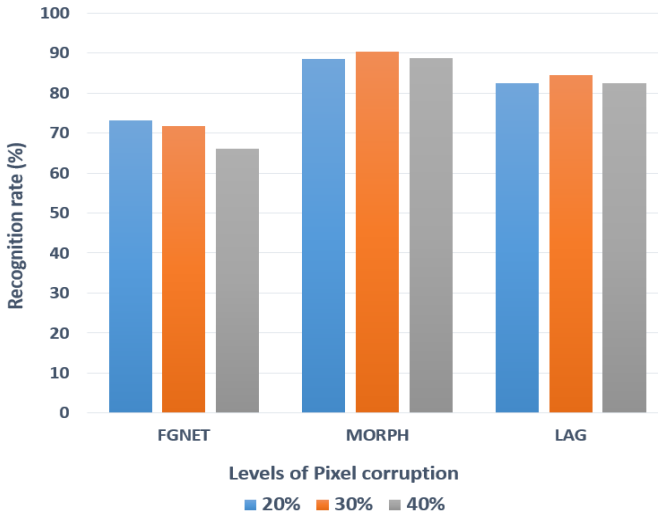


Fig. 5-23. Highest recognition rates of our proposed method using local feature descriptors (DSIFT+LBPD) with optimal feature dimensions.

Table 5-6. Recognition rates under different feature dimensions with 20% of noise on all the three datasets, using local feature descriptors (DSIFT + LBPD).

Feature Dimension	40	60	80	100	120	140
FGNET	60.24%	62.68%	**73.17%**	65.85%	61.46%	64.39%
MORPH	72.5%	79.75%	**88.50%**	84.81%	83.87%	86.68%
LAG	74.33%	76.77%	79%	77.33%	80.55%	**82.44%**

Table 5-7. Recognition rates under different feature dimensions with 30% of noise on all the three datasets, using local feature descriptors (DSIFT + LBPD).

Feature Dimension	40	60	80	100	120	140
FGNET	63.17%	62.93%	**71.70%**	67.56%	61.95%	57.31%
MORPH	84.93%	**90.31%**	81.62%	89.50%	86.12%	88.25%
LAG	77.11%	77.88%	76.33%	75.66%	79.44%	**84.55%**

Table 5-8. Recognition rates under different feature dimensions with 40% of noise on all the three datasets, using local feature descriptors (DSIFT + LBPD).

Feature Dimension	40	60	80	100	120	140
FGNET	62.92%	62.43%	**66.09%**	66.09%	59.26%	59.75%
MORPH	82.87%	83.81%	82.31%	80.56%	87.18%	**88.75%**
LAG	76.44%	75.66%	75.11%	**81.77%**	76.55%	80.77%

5.7 Conclusions

This Chapter presents a novel deep low-rank feature learning and encoding method for cross-age face recognition, when both training and testing images are corrupted by noise. Our method learns discriminative low-rank features by introducing a manifold-constrained low-rank decomposition algorithm. This not only recovers the original image from its corrupted version, but also preserves the local structural information of the data samples. To make the features discriminative in terms of age progression, a locality-based feature-encoding framework is proposed, which encodes low-rank gallery and query image's features using the learned codebook. The encoded features are then fed to nearest neighbor classifier to do face recognition. Furthermore, we also consider the periocular region of a human face instead of a whole face image to do recognition using our proposed framework. Our experimental results on three challenging face-aging datasets demonstrates the superiority of our proposed method.

Chapter 6 Conclusions and Future Research Direction

In this thesis, we first introduce the major theme of our research work by reviewing some of the existing challenges in face-recognition research. Our motivation behind the research is briefly discussed. From Chapter 3 to Chapter 5, we proposed and described complete frameworks for solving the low-resolution and age-invariant face-recognition problems. Furthermore, we take into account the problem of noise variations in face images, which heavily degrade the performance of face-recognition algorithms. In Chapter 3, we proposed an effective approach to solving the low-resolution face-recognition problem, based on sparse discriminant low-rank features. In Chapter 4, a deep-feature encoding-based discriminative model was presented to solve the aging face-recognition problem. In Chapter 5, another age-invariant recognition framework was proposed, which not only can recognize human faces with a large age gap, but also tackle severe noise variations by using deep low-rank feature learning and encoding. In this research, both handcrafted and deep-learning features have been investigated and employed for face recognition. We found that by encoding the features based on sparse representation and low-rank presentation, their discriminant ability can be greatly enhanced.

In this final chapter, we conclude our research findings along with the major contributions. Moreover, future research work will also be discussed.

6.1 Conclusions of our research findings

In our research work, we have attempted to solve three major challenges of face recognition. These include low-resolution face recognition, age-invariant face recognition, and noisy face recognition.

To solve the low-resolution face-recognition problem, an effective method based on sparse coding and low-rank features was proposed in Chapter 3. The proposed method includes the decomposition of extracted local features (Gabor wavelets, and LBPD) into a low-rank feature matrix and a sparse error matrix. The learned low-rank part is then used to learn a common discriminative feature subspace using our proposed sparse coding-based objective function. Experimental results on four challenging face datasets demonstrate

the superiority of this proposed method, which provides a high recognition rate for very low-resolution face images, even of the size 8×8.

To solve the aging face-recognition problem in Chapter 4, we first utilize a pre-trained deep-CNN model (AlexNet) to extract high-level deep features from face images. To make the features more discriminative in terms of age progression, we proposed to learn a codebook from training data, which is then used to encode the gallery and query face image's features, with locality information. To learn an age-discriminative codebook, we first divide training images of each subject into two groups, with a large age gap. After that, the features of these two age group's images are fused using canonical correlation analysis. Finally, these fused features of the training set are used to learn a codebook. The proposed feature-encoding method provides closed form solutions in both the sparse-coding and codebook updating stages. Experimental results on three challenging face-aging datasets demonstrate the superiority of this method. Furthermore, our method does not require any age labels for the recognition purpose.

Another framework for aging face recognition was proposed in Chapter 5, for solving the problem in the presence of significant noise variations. To solve this problem, we proposed a new manifold-constrained low-rank approximation algorithm, which not only recovers the underlying identity information from corrupted face images, but also preserves the local structure of the data samples. The learned low-rank features are then encoded using our proposed feature-encoding framework based on the exponential locality information. Encoded features are then fed to the nearest neighbor classifier for recognition. Various studies have shown that the periocular region of a human face goes through little change with age progression, as compared to other facial parts. Therefore, we also investigated the use of the periocular region of a face image for age-invariant face recognition, based on our proposed framework, in Chapter 5. Our method shows high robustness against noise variations, even if around half of the image pixels are corrupted. It is worth noting that our method is the first one to investigate the performance of aging face recognition with noise variations.

6.2 Future Work

This thesis has presented some techniques for effective facial feature analysis and recognition in both constrained and unconstrained environments. However, there still exist some challenges, which need to be addressed in order to develop a generic facial recognition system. In this section, we will discuss some of our future research directions in the field of face recognition. Future research can be conducted in the following areas:

(1) Recognition of low-resolution (LR) face images: According to some recent research, super-resolution (SR) techniques based on deep-learning can be helpful in improving the recognition accuracy of LR face recognition systems. However, current deep-learning models are not optimal for recognizing LR images of size lower than 12×12. Although deep learning-based SR techniques can generate high-resolution photorealistic images from LR face images, they often lose some identity information during the estimation process. Furthermore, identifying resolution robust features is one of the major challenges. In our future work, we will emphasis on some recent deep architectures which can perform super-resolution, and recognition simultaneously by preserving identity information. Another possible approach is to super-resolve the LR features extracted from deep networks, which can boost the recognition rate.

(2) Age-invariant face-recognition problem: deep-learning has achieved superior performance in many machine-learning tasks, due to the availability of a large amount of training data. To further enhance the recognition accuracy on more challenging aging datasets, a large training set with real age labels should be utilized to learn deep features at different ages. Furthermore, we aim to improve our feature-encoding framework by incorporating manifold information in the objective function. The manifold-learning algorithms can learn an effective projection matrix from collected training data, which can improve the recognition accuracy. Furthermore, the utilization of aging information in manifold information can be helpful in reducing the inter-personal variations. The proposed encoding scheme can also be converted into an end-to-end deep learning framework.

This is possible by designing a new layer, which can be embedded to the feature extraction network to realize both feature fusion and codebook learning. Therefore, the optimization of feature representation and feature fusion can be conducted at the same time and maximize the correlated complementary information between the two steps.

(3) Generalization of face recognition systems: A general face recognition system should be able to tackle all kinds of facial variations, such as pose, expression, illumination, occlusion, noise, resolution, etc. According to our experimental results in Chapter 3, one of the recently proposed deep-CNN models [49], known as SphereFace, cannot perform well when image resolution becomes lower than 20×20. Furthermore, it was argued in [190] that deep models give the worst performance in the presence of noise variations. In our future work, we aim to combine the techniques proposed in this thesis to develop a general face-recognition system, which is robust against all the above-mentioned variations. Regarding the low-rank feature-learning algorithm proposed in Chapter 5, we aim to incorporate sparse-coding information in the objective function. As discussed in Chapter 3, our proposed sparse-coding-based objective function can preserve the local structural information of data samples in the low-dimensional feature subspace, without searching for k-nearest neighbors. This can provide better reconstruction of data samples from its corrupted version, which can be helpful in tackling noise variations.

References

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

[2] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037-2041, Dec. 2006.

[3] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE. Conf. Comput. Vis. Pattern Recognit.,* 2014, pp. 1891-1898.

[4] S. Liao, A.K. Jain, and S.Z. Li, "Partial face recognition: Alignment-free approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1193-1205, May. 2013.

[5] C. Ding, C. Xu, and D. Tao, "Multi-Task Pose-Invariant Face Recognition," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 980-993, Mar. 2015.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2015, pp. 815-823.

[7] J. Wright, A. Y. Yang, and A. Ganesh, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210-227, 2008.

[8] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning Discriminant Face Descriptor," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 36, no. 2, pp. 289-302, Feb. 2014.

[9] W.W. Zou, and P.C. Yuen, "Very low-resolution face recognition problem," *IEEE Trans. Image Process.* vol. 21, no. 1, pp. 327-340, Jan. 2012.

[10] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Sig. Process Lett.*, vol. 17, no. 1, pp. 20-23, Jan. 2010.

[11] Z. Wang, Z. Miao, Q.J. Wu, Y. Wan, and Z. Tang, "Low-resolution face recognition: A review," *The Vis. Comput.*, vol. 30, pp. 359-386, 2014.

[12] D. Deb, L.B-Rowden, and A. K. Jain, “Face Recognition Performance under Aging,” in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit Workshops,* 2017, pp. 1-9.

[13] H. Zhou, and K-M. Lam, “Age-invariant face recognition based on identity inference from appearance age,” *Pattern Recognit.*, vol. 76, pp. 191-202, April 2018.

[14] Z. Li, U. Park, and A. K. Jain, “A discriminative model for age invariant face recognition,” *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 1028–1037, Sep. 2011.

[15] G. B. Huang, M. Ramesh, T. Berg, E. L-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, 2007 (vol. 1, no. 2, p. 3), “Technical Report 07-49, UMass.

[16] K. C. Yow, and R. Cipolla, “Feature-based human face detection,” *Image and Vis. Comput.*, vol. 15, pp. 713-735, 1997.

[17] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, 2004.

[18] C. Liu, “A Bayesian discriminating features method for face detection,” *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 25, no. 6, pp. 725-740, Jun. 2003.

[19] C. Erdem, S. Ulukaya, A. Karaali, and A.T. Erdem, “Combining Haar Feature and skin color based classifiers for face detection,” in *Proc. IEEE Conf. Acoustics. Speech, and Signal Processing (ICASSP)*, 2011, pp. 1497-1500.

[20] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2013, pp. 3476,-3483.

[21] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks*,” IEEE Sig. Process Lett.*, vol. 23, no. 10, pp. 1499-1503, 2016.

[22] S. Milborrow, and F. Nicolls, “Locating facial features with an extended active shape model,” in *Proc. European Conf. Comput. Vis (ECCV)*, 2008, pp. 504-513.

[23] X. Zhu, and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879-2886.

[24] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681-685, Jun. 2001.

[25] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image. Understanding*, vol. 61, pp. 38-59, 1995.

[26] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 7, pp. 711-720, 1997.

[27] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328-340, 2005.

[28] X. He and P. Niyogi, "Locality preserving projections, in: *Proc. Conference on Neural Information Processing Systems (NIPS)*, 2004, pp. 153.

[29] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[30] S.T. Roweis, and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, 2000, pp. 2323-2326.

[31] W.S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401-419, Dec. 1952.

[32] M. Belkin, and P. Niyogi, "Laplacian Eigen maps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[33] L. Wiskott, J-M. Fellous, N. Kruger, and C. V. D. Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, 1997.

[34] B. Kepenekci, F. B. Tek, and G.B. Akar, "Occluded face recognition based on Gabor wavelets," in *Proc. IEEE Conf. Image. Process (ICIP)*, 2002, pp. 293-296.

[35] T.S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 18, no. 10, pp. 959-971, Oct. 1996.

[36] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2005, pp. 1-8.

[37] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no.2, pp. 91-110, 2004.

[38] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput*., vol. C-23, pp. 88–93, Jan. 1974.

[39] A. Vedaldi, and B. Fulkerson, Vlfeat: An open and portable library of computer vision algorithms," in Proc. Int. Conf. Multimedia. 2010, pp. 1469-1472. Available: www.vlfeat.org/

[40] X. Tan, and B. Triggs, "Fusing Gabor and LBP Feature sets for kernel-based Face Recognition," in *Proc. International workshop analysis and modeling of faces and gestures*, 2007, pp. 235-249.

[41] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbour classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[42] T. Joachims, "Text categorization with support vector machine: Learning with many relevant features," in *European Conf. Machine Learning*, 1998.

[43] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, 2010.

[44] J.R. Quinlan, "Induction of decision trees," Machine Learning vol. 1, 1986, pp. 81–106.

[45] P.M. Domingos, and Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classifier," In *Proc. International Conference on Machine Learning (ICML)*, 1996, pp. 105–112.

[46] N. Friedman, D. Geiger, and M. Goldsszmidt, "Bayesian network classifiers," *Journal of Machine learning*, vol. 29, pp. 131-163, 1997.

[47] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron," in *Proc. Third IEEE Conf. Face and Gesture Recognition*, pp. 454-459, Apr. 1998.

[48] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in: *Proc. British. Mach. Vis. Conf. (BMVC)*, 2015, pp. 6.

[49] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," in: *Proc. IEEE Conf. Comput. Vis Pattern Recognit*, 2017.

[50] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European conference on 674 computer vision, 2016, pp. 499-515.

[51] J. Hu, J. Lu, and Y-P. Tan, "Discriminative Deep Metric learning for Face verification in the wild," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2014.

[52] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deep Face: Closing the gap to human-level performance in face verification," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2014.

[53] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," In *Proc. IEEE Conf. Comput. Vis. (ICCV)*, 2013.

[54] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2015.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual learning for Image Recognition," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016.

[56] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. M-Forster, and T. Vetter, "Empirically analysing the effect of dataset biases on deep face recognition systems," *arXiv preprint arXiv: 1712.01619*, 2017.

[57] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," arXiv preprint arXiv: 1411.7923, 2014.

[58] I. K-Shlizerman, S.M. Seitz, D. Miller, and E. Brossard, "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016.

[59] M. Wang, and W. Deng, "Deep Face Recognition: A Survey," arXiv: 1804.06655v4, Jun, 2018.

[60] S. Baker, and T. Kanade, "Hallucinating faces," in: *Proc. IEEE Conf. Auto. Face Gest. Recognit*, 2000, pp. 83-88.

[61] S. Baker, and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 24, no.24, no. 9, pp.1167-1183, 2002.

[62] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861-2873, 2010.

[63] X. Wang, and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Sys. Man. Cyber*, vol. 35, no. 3, pp. 425-434, 2005.

[64] G. Qiu, "A progressively predictive image pyramid for efficient lossless coding," *IEEE Trans. Image Process.*, vol. 8, no. 1, pp. 109–115, 1999.

[65] G. Qiu, "Interresolution look-up table for improved spatial magnification of image," *Journal of Visual Communication and Image Representation*, vol. 11, no. 4, pp. 360–373, 2000.

[66] Y. Hu, K.-M. Lam, G. Qiu, and T. Shen, "From local pixel structure to global image super-resolution: A new face hallucination framework," *IEEE Trans. Image Process.*, vol. 20, no. 2, pp.433–445, 2011.

[67] C.-X. Ren, D.-Q. Dai, and H. Yan, "Coupled kernel embedding for low resolution face image recognition," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3770-3783, 2012.

[68] P.H. Hennings-Yeomans, S. Baker, and B. V. K. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2008, pp. 1-8.

[69] H. Huang, H. He, "Super-Resolution Method for Face Recognition using Nonlinear Mappings on Coherent Features," *IEEE Trans. Neur. Networks*, vol. 22, no. 1, pp. 121-130, 2011.

[70] M. Jian and K.–M. Lam, "Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition," *IEEE Trans. Circuits Syst. Video Technol*., vol. 25, no. 11, pp. 1761–1772, Nov. 2015.

[71] K.-H. Pong, and K.-M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognit,* vol. 47, no. 2, pp. 556-567, 2014.

[72] C. Zhou, Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Low-resolution face recognition via simultaneous discriminant analysis," in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–6.

[73] S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution Face images," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 34, no. 10, pp. 2019-2030, 2012.

[74] S. Siena, V.N. Bodetti, and B.V. Kumar, "Coupled marginal fisher analysis for low-resolution face recognition," in: *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 240-249.

[75] J. Shi, and C. Qi, "From Local Geometry to Global Structure: Learning Latent Subspace for Low-resolution Face Image Recognition," *IEEE Signal Process Lett*, vol. 22, no. 5, pp. 554-558, 2015.

[76] J. Zhang, Z. Guo, X. Li, and Y. Chen, "Large Margin Coupled Mapping for Low Resolution Face Recognition," in: *Proc. Pacific Rim International conference on Artificial Intelligence*, 2016, pp. 661-672.

[77] Z. Wang, W. Yang, and X. Ben, "Low-resolution degradation face recognition over long distance based on CCA," *Neural Comput. Applicat*, vol. 26, no. 7, pp. 1645-1652, 2015.

[78] J. Jiang, R. Hu, Z. Wang, and Z. Cai, "CDMMA: Coupled discriminant multi-manifold analysis for matching low-resolution face images," *Signal Process.*, vol. 124, pp. 162-172, 2016.

[79] Y. Chu, T. Ahmad, G. Bebis, and L. Zhao, "Low-resolution face recognition with single image per person," *Signal Process*, vol. 141, pp.144-157, 2017.

[80] X. Xing, and K. Wang, "Couple manifold discriminant analysis with bipartite graph embedding for low-resolution face recognition," *Signal Process*, vol. 125, pp. 329-335, 2016.

[81] F. Yang, W. Yang, R. Gao, and Q. Liao, "Discriminative multidimensional scaling for Low-resolution face recognition," *IEEE Signal Process Lett*, vol. 25, no. 3, pp. 388-392, 2018.

[82] Y. Fu, and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 578–584, Jun. 2008.

[83] X. Geng, Z.-H. Zhou, and K. Smith-Mile, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, 2007.

[84] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[85] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2009, pp. 112-119.

[86] Y. H. Kwon and N. D. V. Lobo, "Age classification from facial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 1999, pp. 762–767.

[87] A. Lanitis, C. Draganova, and C. Christodoulou, "Comparing different classifiers for automatic age estimation," *IEEE Trans. Syst., Man, Cybern. B, Cybern*, vol. 34, no. 1, pp. 621–628, Feb. 2004.

[88] N. Ramanathan and R. Chellappa, "Face verification across age progression," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3349–3361, Nov. 2006.

[89] J.-X. Du, C.-M. Zhai, and Y.-Q. Ye, "Face aging simulation and recognition based on NMF algorithm with sparseness constraints," *Neurocomputing*, vol. 116, pp. 250–259, Sep. 2013.

[90] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):442–455, 2002.

[91] U. Park, Y. Tong, and A. K. Jain, "Age-invariant face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 947–954, May 2010.

[92] J. Suo, X. Chen, S. Shan, and W. Gao, "Learning long term face aging patterns from partially dense aging databases," in *Proc. 12th ICCV*, 2009, pp. 622–629.

[93] J. Suo, S.-C. Zhu, S. Shan, and X. Chen, "A compositional and dynamic model for face aging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 385–401, Mar. 2010.

[94] N. Tsumura *et al.*, "Image-based skin color and texture analysis/ synthesis by extracting hemoglobin and melanin information in the skin," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 770–779, 2003.

[95] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2006, pp. 387-394.

[96] J. Wang, Y. Shang, G. Su, and X. Lin, "Age simulation for face recognition," in *Proc. 18th ICPR*, vol. 3. 2006, pp. 913–916.

[97] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, "Face verification across age progression using discriminative methods," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 1, pp. 82–91, Mar. 2010.

[98] G. Guo, G.Mu, and K. Ricanek, "Cross-age face recognition on a very large database: The performance versus age intervals and improvement using soft biometric traits," in *Proc Int. Conf. Pattern Recognit.*, 2010, pp. 3392–3395.

[99] B. Klare and A. K. Jain, "Face recognition across time lapse: On learning feature subspaces," in *Proc. IEEE Conf. Joint Biometrics.*, Washington, DC, USA, Oct. 2011, pp. 1–8.

[100] C. Otto, H. Han, and A. K. Jain, "How does aging affect facial components?" in *Proc. ECCV*, Florence, Italy, Oct. 2012, pp. 189–198.

[101] D. Gong, Z. Li, and D. Tao, "A Maximum Entropy Feature Descriptor for age-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2015, pp. 5289-5297.

[102] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *Proc. ICCV*, 2013, pp. 2872–2879.

[103] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. ECCV*, 2014, pp. 768–783.

[104] L. Du and H. Ling, "Cross-age face verification by coordinating with cross-face age verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2015, pp. 2329-2338.

[105] Z. Li, D. Gong, X. Li, and D. Tao, "Aging face recognition: A Hierarchical learning model based on local patterns selection, *IEEE Trans. Image Process*., vol. 25, no. 5, pp. 2146-2154, May 2016.

[106] Y. Wen, Z. Li, and Y. Qiao, "Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2016, pp. 4893-4901.

[107] T. Zheng, W. Deng, and J. Hu, "Age Estimation Guided Convolutional Neural Network for Age-Invariant Face Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 503-511.

[108] H. Li, H. Hu, and C. Yip, "Age-Related Factor Guided Joint Task Modeling Convolutional Neural Network for Cross-Age Face Recognition," *IEEE Trans. Inf. Forensics Security,* vol.13, no. 9, Sep. 2018.

[109] C. Xu, Q. Liu, and M. Ye, "Age invariant face recognition and retrieval by coupled auto-encoder networks," *Neurocomputing*, vol. 222, pp. 62-71, 2017.

[110] Y. Li, G. Wang, L. Nie, Q. Wang, and W. Tan, "Distance metric optimization driven convolutional neural network for age invariant face recognition," *Pattern Recognit.*, vol. 75, pp. 51-62, 2018.

[111] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang, "Orthogonal Deep Features Decomposition for Age-invariant Face Recognition," *arXiv:1810.07599*, 2018.

[112] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of key points," In Workshop on Statistical Learning in Computer Vision, *ECCV*, pp. 1–22, 2004.

[113] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial pyramid matching for recognizing Natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2006.

[114] J. Sivic, and A. Zisserman, Video Google: A text retrieval approach to object matching in videos," In *Proc. IEEE Conf. Comput. Vis. (ICCV)*, 2003.

[115] J.A. Hartigan, and M.A. Wong, "Algorithm AS136: A k-means Clustering Algorithm," *Applied Statistics*, vol. 28, pp. 100-108, 1979.

[116] S.P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Infor. Theory*, vol. 28, pp. 129-137, 1982.

[117] C. Rasmussen, "The Infinite Gaussian Mixture Model," Advances in Neural Information Processing Systems, vol. 12, pp. 554-560, 2000.

[118] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/ discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712-727, April 2008.

[119] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, "Unifying discriminative visual codebook generation with classifier training for object category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2008, pp. 1-8.

[120] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2009, pp. 1794-1801.

[121] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Neural Information Processing Systems (NIPS)*, 2009, pp. 2223-2231.

[122] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2010, pp. 3360-3367.

[123] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," In *Proc. Eur. Conf. Comput. Vis. (ECCV*), 2010.

[124] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Largescale image retrieval with compressed Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2010.

[125] R. G. Cinbis, J. Verbeek, and C. Schmid, "Image categorization using Fisher kernels of non-iid image models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2012.

[126] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," 2014, http://arxiv.org/abs/1403.6382.

[127] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., 2013, pp. 3025-3032.

[128] S. Yan, X. Xu, D. Xu, S. Lin, and X. Li, "Beyond spatial pyramids: A new feature extraction framework with dense spatial sampling for image classification," in *Proc. Eur. Conf. Comp. Vis*., 2012, pp. 473–487.

[129] L. Bo, X. Ren, and D. Fox, "Hierarchical matching pursuit for image classification: Architecture and fast algorithms," in *Proc. Adv. Neural Inf. Process. Syst*., 2011, pp. 2115–2123.

[130] L. Liu, C. Shen, L. Wang, A. van den Hengel, and C. Wang, "Encoding high dimensional local features by sparse coding based Fisher vectors," in *Proc. Advances in Neur. Infor. Process Syst.,* 2014.

[131] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Towards a practical face recognition system: Robust Alignment and Illumination by Sparse Representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372-386, Feb. 2012.

[132] L. Zhang, M. Yang, and X. Feng, "Sparse representation or Collaborative representation: Which helps Face Recognition," in: *Proc. IEEE Conf. Comput. Vis. (ICCV),* 2011.

[133] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *Proc. Eur. Conf. Comp. Vis*., 2010.

[134] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," *in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 625–632.

[135] F. De la Torre and M. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, no. 1, pp. 117–142, 2003.

[136] Q. Ke and T. Kanade, "Robust *L*1 norm factorization in the presence of outliers and missing data by alternative convex programming," *Ph.D. dissertation, Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA,* USA, 2005.

[137] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM,* vol. 58, no. 3, article no. 11, 2011.

[138] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using Markov random fields," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 1050–1057.

[139] C.-F. Chen, C.-P. Wei, and Y.-C. F. Wang, "Low-rank matrix recovery with structural incoherence for robust face recognition," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit., CVPR*, Jun. 2012, pp. 2618–2625.

[140] C.-P. Wei, C.-F. Chen, and Y.-C. F. Wang, "Robust Face Recognition with structurally incoherent low-rank matrix decomposition," *IEEE Trans. Image Process,* vol. 23, no. 8, pp. 3294-3307, Aug. 2014.

[141] X-Y. Jing, F. Wu, X. Zhu, X. Dong, F. Ma, and Z. Li, "Multi-spectral low-rank structured dictionary learning for face recognition," *Pattern Recognition*, vol. 59, pp. 14-25, 2016.

[142] F. Wu, X-Y. Jing, X. You, D. Yue, R. Hu, and J-Y. Yang, "Multi-view low-rank dictionary learning for image classification," *Pattern Recognition,* vol. 50, pp. 143-154, Feb. 2016.

[143] F. J. Xu, K. Luu, M. Savvides, T.D. Bui, and C.Y. Suen, Investigating Age invariant face recognition based on Periocular Biometrics, in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–7.

[144] Facial Image Processing and Analysis (FIPA). *FG-NET Aging Database*. [Online]. Available: http://fipa.cs.kit.edu/433.php#Downloads.

[145] J. Merkow, B. Jou, M. Savvides, "An exploration of gender identification using only the periocular region," in *Proc. Int. Conf. Biometrics of the Theory Applications and Systems* (BTAS), 2010, pp. 1–5 .

[146] Y. Dong, D. Woodard, "Eyebrow shape-based features for biometric recognition and gender classification: a feasibility study," in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–8.

[147] F. Smeraldi, and J. Bigun, "Retinal vision applied to facial features detection and face authentication," *Pattern Recognit. Lett.* , vol. 23, pp. 463-475, 2002.

[148] P.E. Miller, J.R. Lyle, S.J. Pundlik, and D.L. Woodard , "Performance evaluation of local appearance based periocular recognition," in *Proc. Int. Conf. Biometrics of the Theory Applications and Systems* (BTAS), 2010.

[149] U. Park, R.R. Jillela, A. Ross, and A.K. Jain, "Periocular biometrics in the visible spectrum," *IEEE Trans. Inf. Forensics Security*, vol. 6, no.1, pp. 96–106, Mar. 2011.

[150] R. Jillela, and A. Ross, "Mitigating effects of plastic surgery: Fusing face and ocular bio- metrics," in *Proc. Int. Conf. Biometrics of the Theory Applications and Systems* (BTAS), 2012, pp. 402–411.

[151] G. Mahalingam , K. Ricanek , and A. Albert , "Investigating the periocular-based face recognition across gender transformation," *IEEE Trans. Inf. Forensics Security*, vol. 9, pp.2180–2192, 2014.

[152] S. Zhang, and X. Zhao, "Locality-sensitive kernel sparse representation classification for face recognition," *Journal of Visual Communication and Image Representation*, vol. 25, pp.1878-1885, 2014.

[153] X. Hong, G. Zhao, M. Pietikainen, and X. Chen, "Combining LBP Difference and Feature correlation for texture description," *IEEE Trans. Image Process.*, vol. 23, no. 6, 2014.

[154] X. He, D. Cai, S. Yan, and H. J. Zhang, "Neighbourhood preserving embedding," in: *Proc. IEEE Conf. Comput. Vis. (ICCV)*, 2005, pp. 1208-1213.

[155] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[156] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Trans. Image Process*., vol. 19, no. 6, pp. 1657–1663, 2010.

[157] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBPV) with global matching," *Pattern Recognit.*, vol. 43, no. 3, pp. 706–719, 2010.

[158] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen, "Rotation-invariant image and video description with local binary pattern features," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1465-1477, 2012.

[159] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Communications on Pure and Applied Mathematics*, vol. 30, no. 5, pp. 509-541, 1977.

[160] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[161] H. Bay, A. Ess, T. Tuytclaars, and L.-V. Gool, "Speeded-Up Robust Features," *Computer vision and image understanding*, vol. 110, pp. 346-359, 2008.

[162] N. Dalal, and B. Triggs, "Histograms of Oriented Gradients for human detection," in *Proc. IEEE Conf., Comput. Vis. Pattern Recognit., CVPR*, 2005.

[163] T.-S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 10, pp. 959-971, 1996.

[164] X. Tan, and B. Triggs, "Enhanced local texture feature sets for face recognition under different lighting conditions, *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635-1650, 2010.

[165] K. Okajima, "Two-dimensional Gabor-type receptive field as derived by mutual information maximization," *Neural Networks*, vol. 11, no. 3, pp. 441-447, 1998.

[166] L. Zhang, and C. Ma, Low-Rank, "Sparse matrix decomposition and group sparse coding for image classification," in *Proc. IEEE Conf. Image. Process (ICIP)*, 2012, pp. 669-672.

[167] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *UIUC 720 Technical Report UILU-ENG-09-2215, Tech. Rep.*, 2009.

[168] V.D. Maaten, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2431-2456, 2008.

[169] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807-813, 2010.

[170] P. J. Philips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 22, no. 10, pp. 1090-1104, 2000.

[171] R. Chellappa, J. Ni, V-M. Patel, "Remote identification of faces: Problems, prospects and progress, *Pattern. Rec. Lett*., vol. 33, no. 14, pp. 1849-1859, 2012.

[172] M. Chevalier, N. Thome, M. Cord, J. Fournier, G. Henaff, and E. Dusch, "LR-CNN for fine-grained classification with varying resolution," in *Proc. IEEE Conf. Image. Process (ICIP)*, 2015, pp. 3101-3105.

[173] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," *in Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2011, pp. 529–534.

[174] J. Kannala, and E. Rahtu, "BSIF: binarized statistical image features," in *Proc. IEEE Conf. Pattern. Recog (ICPR),* 2012.

[175] J. Chen, V.M. Patel, L. Liu, V. Kellokumpu, G. Zhao, M. Pietikainen, and R. Chellappa, "Robust local features for remote face recognition," *Image and Vision Computing*, vol. 64, pp. 34-46, 2017.

[176] G.E. Hinton, and R.R. Salukhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[177] M.S. Shakeel, and K.-M. Lam, "Recognition of Low Resolution Face Images using Sparse Coding of Local Features," in: *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA),* 2016, pp. 1-5.

[178] M. Bereta, P. Karczmarek, W. Pedrycz, and M. Reformat, "Local descriptors in application to the aging problem in face recognition," *Pattern Recognition*, vol. 46, no. 10, pp. 2634-2646, 2013.

[179] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. Neural Information Processing Systems (NIPS)*, 2012, pp. 1106-1114.

[180] A. Elgammal, R. Duraiswami, and L. Davis, "Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 11, pp. 1499–1504, 2003.

[181] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327-1336, 2011.

[182] K. Ricanek, Jr., and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th FGR*, 2006, pp. 341–345.

[183] S. Bianco, "Large Age-gap Face verification by Feature Injection in Deep Networks," *Pattern Recognition letters*, vol. 90, pp. 36-42, 2017.

[184] H.V. Nguyen, and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Comput. Vis. (ACCV)*, Springer, 2010, pp. 709-720.

[185] L. Wolf, T. Hassner, and Y. Taigman, "The one-shot similarity kernel," in *Proc. 12th ICCV*, 2009, pp. 897–902.

[186] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. ECCV*, 2012, pp. 566–579.

[187] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proc. ICCV*, 2013, pp. 2408–2415.

[188] J. L. Bentley, "Multidimensional Binary Search Trees Used for Associative Searching," *Comm. ACM*, vol.18, pp. 509-517, 1975.

[189] J. Pang, L. Qin, C. Zhang, W. Zhang, Q. Huang, and B. Yin, "Local Laplacian coding from Theoretical Analysis of Local Coding Schemes for Locally Linear Classification," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2937-2947, 2015.

[190] S. Dodge, and L. Karam, "Understanding How Image Quality Affects Deep Neural Networks," in *Proc. International Conf. on Quality of Multimedia Experience*, 2016, pp. 1-6.

[191] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. *International conference on Learning Representation (ICLR)*, 2015.

[192] E. Candes, and B. Recht, "Exact low-rank matrix completion via convex optimization," in: *Proc. Allerton*, 2008.

[193] A. Ganesh, Z. Lin, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast algorithms for recovering a corrupted low-rank matrix, in *Proc. IEEE Int. Workshops on Computational Advances in Multi-sensor Adaptive Processing (CAMSAP)*, 2009, pp. 213-216.

[194] V. Chandrasekaran, S. Sanghavi, P.A. Parillo, and A.S. Willsky, "Rank-sparsity incoherence for Matrix decomposition," *SIAM J. Optim*, vol. 21, no. 2, pp. 572-596, 2011.

[195] X. Jiang, and J. Lai, "Sparse and dense hybrid representation via dictionary decomposition for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1067-1079, 2015.

[196] V.M. Patel, Y-C. Chen, R. Chellappa, and P. J. Phillips, "Dictionaries for image and video-based face recognition," *J. Opt. Soc. Am*, vol. 31, no. 5, pp. 1090-1103, May 2014.

[197] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained Alex Net Architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote sensing*, vol. 9, no. 8, pp. 848.

[198] K. Tang, X. Hou, Z. Shao, and L. Ma, "Deep Feature Selection and Projection for Cross-Age Face Retrieval," in *Proc. Int. Cong. Image. Sig. Proc, Bio Medical Eng. Inform.* 2017.