Robust Face Recognition

Changxing Ding

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

July, 2016

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Changxing Ding
Date: 06/07/2016

I would like to dedicate this thesis to my loving wife and parents.

Acknowledgements

Many people have significantly influenced me - both academically and personally - during my time at University of Technology Sydney (UTS). I would like to take this opportunity to express my sincere gratitude to them for their support during my PhD study.

My deepest gratitude goes to Prof. Dacheng Tao, who has been not just a great supervisor but also a sincere friend. Prof. Tao has taught me how to perform research from scratch and has guided me into academia. He has been a constant source of suggestions and inspiration, and I have benefitted significantly from our discussions and his mentorship. I would not have been able to publish scientific papers in top journals without his insightful instructions, high scientific standards, patient guidance, and consistent encouragement. Also, many thanks to his guidance and suggestions for my future career during the final stage of my PhD study.

I owe special thanks to Prof. Larry S. Davis at the University of Maryland, College Park (UMD) for a number of insightful discussions and suggestions on my first work on face recognition. His careful criticism bridging theory and application broadened my thinking and helped me to improve the work. I would also like to thank Dr. Jonghyun Choi from UMD for his collaboration on experiments and developing the first work. I also thank Dr. Chang Xu, Beijing University, who answered some of my general questions about machine learning and helped me develop the model for pose-invariant face recognition.

I have been so fortunate to work in the Centre for Quantum Computation and Intelligent Systems (QCIS). I appreciate the support of the center director, Prof. Chengqi Zhang, who has led QCIS to be a leading research center where I had the chance to meet and get to know many world-famous

researchers. I also want to express my sincere appreciation to the other professors and staff at QCIS for their generous support: Prof. Mingsheng Ying, Prof. Jie Lu, Prof. Ivor Tsang, Prof. Yi Yang, Dr. Ling Chen, Dr. Peng Zhang, Dr. Guodong Long, and Dr. Jing Jiang. I have really enjoyed the time that I have spent with my brilliant colleagues: Dr. Wei Bian, Dr. Tianyi Zhou, Dr. Jun Li, Dr. Zhibin Hong, Dr. Meng Fang, Mingming Gong, Tongliang Liu, Maoying Qiao, Chen Gong, Qiang Li, Ruxin Wang, Shaoli Huang, Zhe Xu, Xiyu Yu, Hao Xiong, Yali Du, and Jiankang Deng. I also wish to express my appreciation to other friends I met in Sydney: Prof. Weifeng Liu, Prof. Lianwen Jin, Prof. Bo Du, Dr. Naiyang Guan, Dr. Nannan Wang, Dr. Chunyang Liu, Dr. Xianye Ben, Dr. Wankou Yang, and Dr. Shigang Liu. Their companionship has always been a source of strength in my daily life. I also owe thanks to other friends at QCIS who volunteered to have their photos included in my face database.

Thanks go to Dr. Matthew Gaston, cluster administrator, and other faculty staff. Dr. Gaston patiently answered my numerous questions and spent a lot of time setting up and managing our servers. My research could not have been conducted successfully without his hard work.

It is my honor to acknowledge sponsorship from the China Scholarship Council and UTS. Their generosity has kept me free from funding concerns. I owe a big thanks to QCIS, who generously provided me with extra scholarship funds and supported me at CVPR conferences over two successive years. I am greatly indebted to Prof. Huijun Gao and Prof. Ligang Wu, who supervised my Masters studies at the Harbin Institute of Technology. I would not have had the chance to pursue my PhD degree under Prof. Tao's supervision without their guidance and generous support.

Last but not least, I want to express my special thanks to my family: my wife Feifei Liu, my parents, and my grandparents for their selfless love and unwavering support during my PhD study. I dedicate this dissertation to them.

Abstract

Face recognition is one of the most important and promising biometric techniques. In face recognition, a similarity score is automatically calculated between face images to further decide their identity. Due to its non-invasive characteristics and ease of use, it has shown great potential in many real-world applications, e.g., video surveillance, access control systems, forensics and security, and social networks. This thesis addresses key challenges inherent in real-world face recognition systems including pose and illumination variations, occlusion, and image blur. To tackle these challenges, a series of robust face recognition algorithms are proposed. These can be summarized as follows:

In Chapter 2, we present a novel, manually designed face image descriptor named "Dual-Cross Patterns" (DCP). DCP efficiently encodes the seconder-order statistics of facial textures in the most informative directions within a face image. It proves to be more descriptive and discriminative than previous descriptors. We further extend DCP into a comprehensive face representation scheme named "Multi-Directional Multi-Level Dual-Cross Patterns" (MDML-DCPs). MDML-DCPs efficiently encodes the invariant characteristics of a face image from multiple levels into patterns that are highly discriminative of inter-personal differences but robust to intra-personal variations. MDML-DCPs achieves the best performance on the challenging FERET, FRGC 2.0, CAS-PEAL-R1, and LFW databases.

In Chapter 3, we develop a deep learning-based face image descriptor named "Multimodal Deep Face Representation" (MM-DFR) to automatically learn face representations from multimodal image data. In brief, convolutional neural networks (CNNs) are designed to extract

complementary information from the original holistic face image, the frontal pose image rendered by 3D modeling, and uniformly sampled image patches. The recognition ability of each CNN is optimized by carefully integrating a number of published or newly developed tricks. A feature level fusion approach using stacked auto-encoders is designed to fuse the features extracted from the set of CNNs, which is advantageous for non-linear dimension reduction. MM-DFR achieves over 99% recognition rate on LFW using publicly available training data.

In Chapter 4, based on our research on handcrafted face image descriptors, we propose a powerful pose-invariant face recognition (PIFR) framework capable of handling the full range of pose variations within $\pm 90^{\circ}$ of yaw. The framework has two parts: the first is Patch-based Partial Representation (PBPR), and the second is Multi-task Feature Transformation Learning (MtFTL). PBPR transforms the original PIFR problem into a partial frontal face recognition problem. A robust patch-based face representation scheme is developed to represent the synthesized partial frontal faces. For each patch, a transformation dictionary is learnt under the MtFTL scheme. The transformation dictionary transforms the features of different poses into a discriminative subspace in which face matching is performed. The PBPR-MtFTL framework outperforms previous state-of-the-art PIFR methods on the FERET, CMU-PIE, and Multi-PIE databases.

In Chapter 5, based on our research on deep learning-based face image descriptors, we design a novel framework named Trunk-Branch Ensemble CNN (TBE-CNN) to handle challenges in video-based face recognition (VFR) under surveillance circumstances. Three major challenges are considered: image blur, occlusion, and pose variation. First, to learn blur-robust face representations, we artificially blur training data composed of clear still images to account for a shortfall in real-world video training data. Second, to enhance the robustness of CNN features to pose variations and occlusion, we propose the TBE-CNN architecture, which efficiently extracts complementary information from holistic face images and patches cropped around facial components. Third, to further promote

the discriminative power of the representations learnt by TBE-CNN, we propose an improved triplet loss function. With the proposed techniques, TBE-CNN achieves state-of-the-art performance on three popular video face databases: PaSC, COX Face, and YouTube Faces.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Rapid technological developments and the popularity of camera devices have fuelled a growing demand for personal identification by cameras and computers. Inherent and unique human characteristics, for example faces, fingerprints, iris patterns, and gait, are employed for more secure and convenient recognition. Recognition technology based on these human appearance characteristics is known as biometrics.

Of all the biometric techniques, face recognition has received the most attention. The aim of face recognition is to automatically recognize human identities from face images using computer vision algorithms. Compared with other biometric techniques, face recognition has two key advantages. First, it is a passive biometric technology that can recognize non-cooperative subjects, and this non-invasiveness is particularly important for surveillance circumstances where recognition needs to be performed by hidden cameras. In contrast, fingerprint and iris recognition require cooperative subjects and is invasive. Second, face recognition requires only ordinary camera devices; therefore, hardware costs are low. In comparison, fingerprint and iris recognition rely on specially designed imaging hardware (and therefore expensive). Therefore, face recognition has become the most popular and convenient biometric technique in practice and has aroused interest by both academia and industry.

A face recognition system typically has five components: face image acquisition, face and facial landmark detection, face normalization and pre-processing, facial

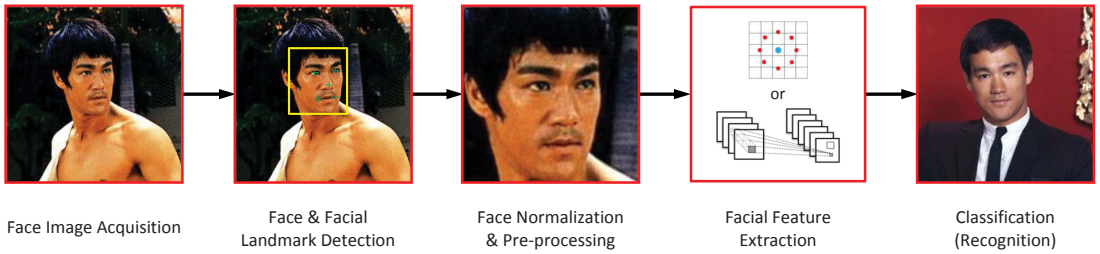


Figure 1.1: A typical face recognition system pipeline. The image is selected from the CASIA-WebFace database [Yi et al., 2014].

feature extraction, and classification (Fig. 1.1).

- **Face image acquisition**: An imaging device captures an image including at least one face.
- **Face & facial landmark detection**: A face detection algorithm detects the position and size of the face. The coordinates of facial feature points, e.g., eyes, nose tip, and mouth corners, are also detected using face alignment algorithms. For video-based face recognition, face tracking may also be applied.
- **Face normalization & pre-processing**: The face image is geometrically normalized by affine/similarity transformations according to the coordinates of the facial feature points, such that facial components in different face images are aligned in the normalized images. More complicated geometric normalization techniques, e.g., 2D or 3D pose normalization [Ding and Tao, 2016], are optionally used to reduce the impact of face pose variations. Photometric normalization maybe applied in some circumstances to suppress illumination variations [Han et al., 2013]. Classical works include Self Quotient Image [Wang et al., 2004], Logarithmic Total Variation Model [Chen et al., 2006], and Difference of Gaussian [Tan and Triggs, 2010]. There are also works that turn the gray-scale image into a set of gradient images, e.g., gradient filters [Ding et al., 2016] and Gabor filters [Xie et al., 2010].
- **Facial feature extraction**: This is the most critical part of a face recognition system. A feature vector is extracted from the normalized face image as its representation. A good face representation should be highly discriminative to

distinguish faces of different subjects and highly robust to rich intra-personal facial appearance variations. A detailed survey of existing facial feature extraction methods can be found in Section 1.2.

- **Face classification**: Classification is based on the similarity score between face representations from different face images. The process usually involves two recognition tasks: face identification and face verification. Face identification refers to determining the identity of the input face by comparing its representation with those of known face images in the database. Face verification refers to determining whether two face images are from the same subject, based on their representations. Representative classification models are summarized in Section 1.3.

Besides its practical significance, face recognition is also a classical pattern recognition problem. Progress in face recognition has provided important inspiration for other pattern recognition topics, making it an active research topic for over 40 years. Progress has been made, and mature face recognition products are now available for certain applications in constrained environments such as access control systems. However, face recognition in many real-world circumstances, e.g., large pose variation [Ding and Tao, 2016], video surveillance [Beveridge et al., 2015], and large-scale face recognition [Kemelmacher-Shlizerman et al., 2015], remains very challenging for two main reasons:

- **Small inter-personal appearance variations**: All human faces are of similar shape and component configuration. Differences in human faces often lie in the detail of facial textures, which face recognition algorithms find difficult to distinguish. In particular, in large-scale face recognition [Kemelmacher-Shlizerman et al., 2015], there is an increasing chance that faces share similar appearances, and thus discriminating between them becomes more difficult.

- **Large intra-personal appearance variations**: Face images from the same subject are often dramatically different. Two factors explain rich intra-personal facial appearance variations. First, the face itself can exhibit variations, e.g., pose [Ding and Tao, 2016], expression [Sim et al., 2003], makeup, aging [Suo et al., 2010], and occlusion [Wright et al., 2009]. Second, imaging conditions,

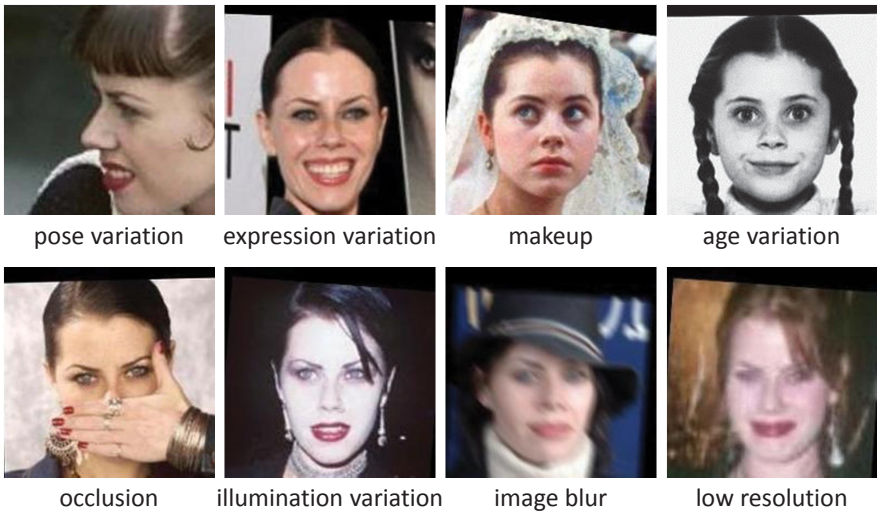


Figure 1.2: The two typical factors that affect facial appearance: variations exhibited by the face itself and variations caused by imaging conditions.

e.g., illumination [Han et al., 2013], image blur [Nishiyama et al., 2011], low resolution, and noise caused by low-cost imaging devices [Beveridge et al., 2015] cause variations. Illustrations of intra-personal facial appearance variations are shown in Fig. 1.2.

To handle these challenging factors, numerous face recognition algorithms have been proposed over the last 40 years. Some representative works and those most relevant to this thesis are reviewed below. For more comprehensive reviews, readers are referred to some good surveys [Ding and Tao, 2016, Tan et al., 2006, Zhao et al., 2003]. Existing works are summarized from the following three perspectives: first, classical approaches to facial feature extraction are reviewed in Section 1.2, because face representation is the most important part of a face recognition algorithm; second, representative classification models are introduced in Section 1.3; and third, existing pose-invariant face recognition (PIFR) algorithms are separately surveyed in Section 1.4, because pose variation is the most challenging factor in face recognition to the extent that PIFR has developed into an independent topic in its own right.

Since publicly available face databases are indispensable to face recognition research, popular face databases are described in Section 1.5. Finally, the contributions of and publications related to this thesis are introduced in Section 1.6.

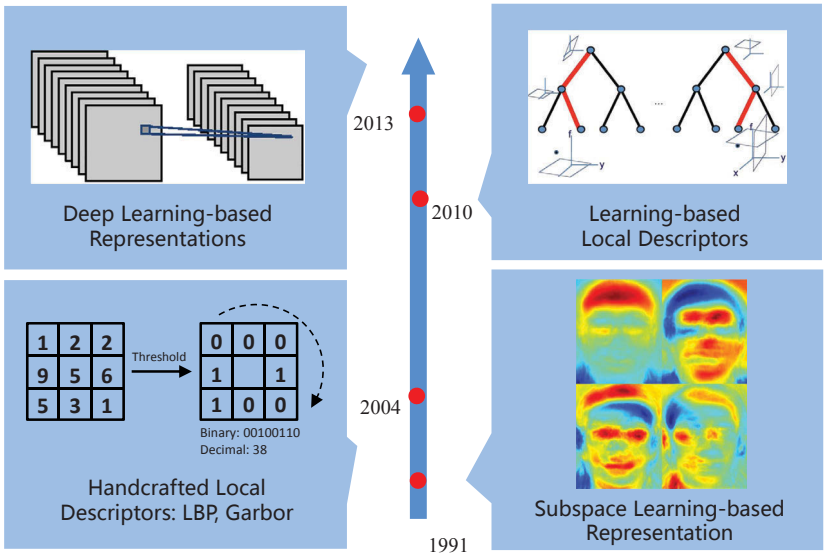


Figure 1.3: The evolution of facial feature extraction methods.

1.2 Facial Feature Extraction

Designing a good facial feature extractor is essential for face recognition systems, because the extracted face representation largely determines system performance. Good face representations are robust to intra-personal appearance variations and discriminative to inter-personal appearance variations. There is a vast literature on facial feature extraction that can be divided into three main categories: subspace learning-based methods, local descriptor-based methods, and deep learning-based methods. The evolution of facial feature extraction methods is illustrated in Fig. 1.3.

1.2.1 Subspace Learning-based Representations

The early representations were subspace learning-based. Face images are usually high dimensional; therefore, redundant information challenges the efficiency and robustness of recognition algorithms. To handle this problem, subspace learning-based approaches project the high-dimensional face image data to a low-dimensional subspace. The projected vector in the low-dimensional subspace forms the face image representation. Since the subspace dimension is usually significantly smaller than that of the face image, the learnt representation is very compact and recognition is more

efficient.

The first subspace learning-based approach for face recognition was the Eigenfaces method [Turk and Pentland, 1991]. Eigenfaces learns a set of subspace bases by Principal Component Analysis (PCA) from unlabeled face image data. Its objective function is based on image reconstruction such that there is minimal energy loss in the learnt subspace. Eigenfaces became one of the classical methods for face recognition, but since Eigenfaces does not make use of label information, the learnt subspace is not sufficiently discriminative for classification.

Inspired by Eigenfaces, many methods have been proposed to effectively utilize label information for discriminative subspace learning. The most representative method is Fisherfaces [Belhumeur et al., 1997], in which a set of subspace bases W are learnt by Linear Discriminant Analysis (LDA), whose objective function maximizes the ratio of between-class scatter matrix S_b to within-class scatter matrix S_w in the low-dimensional subspace:

$$W = \arg\max \frac{|W^T S_b W|}{|W^T S_w W|}. \tag{1.1}$$

$|\cdot|$ stands for the determinant of a matrix. One major limitation of LDA is the "Small Sample Size" (SSS) problem, because the number of training images is usually smaller than the image data dimension; therefore, S_w becomes singular. To address this problem, the authors proposed a two-step method that combines PCA and LDA. In brief, face data is first projected to a PCA subspace to reduce the image data dimension. LDA is then performed in the low-dimensional subspace.

Another representative approach that tackles the SSS problem is Null-space LDA (N-LDA) [Chen et al., 2000], in which the authors suggest that the null-space of S_w contains the most discriminative information. The principle of N-LDA is to choose projection vectors that maximize $W^T S_b W$ with the constraint that $|W^T S_w W|$ is zero. In this way, the criteria in Eq. 1.1 definitely reach a maximum. However, a limitation of N-LDA is that the null-space of S_w becomes smaller with an increase in training data; therefore, information outside the null-space will be lost.

Other popular subspace face recognition methods include Kernelized LDA (KLDA) [Baudat and Anouar, 2000], Two-Dimensional PCA [Yang et al., 2004], and Random Subspace LDA [Wang and Tang, 2006]. Wang *et al.* [Wang and Tang, 2004]

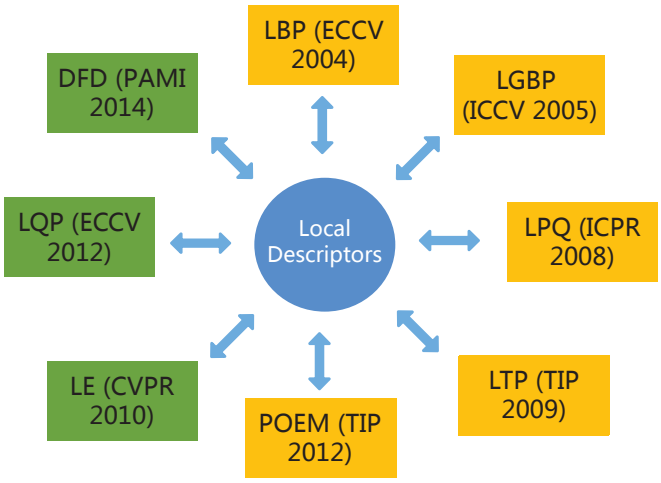


Figure 1.4: The evolution of local descriptors for face image representation. A yellow background represents handcrafted local descriptors and a green background indicates learning-based local descriptors.

unified representative subspace learning-based approaches into a common framework. Subspace learning-based approaches are generally simple and efficient, but they are limited by representing each face in a holistic manner, i.e., each element in the face representation is determined by all the pixels in the image; therefore, subspace learning-based methods suffer when images are occluded or show pose and expression variations [Bian and Tao, 2011]. As a result, subspace-learning models are now seldom employed to extract face representations from raw image pixels. Instead, they are utilized as dimension reduction tools for more effective face image representations [Sun et al., 2014, Xie et al., 2010].

1.2.2 Local Descriptor-based Representations

In contrast to subspace learning-based approaches, local descriptor-based approaches represent a face image by encoding local texture patterns. As each element in the feature vector is determined solely by a local patch, the local descriptor-based representations are more robust to face appearance variations caused by occlusion and pose and expression variations. The evolution of local descriptors is illustrated in Fig. 1.4. A local descriptor is usually designed in three steps: image filtering, local sampling, and pattern encoding, of which the last two steps are the most

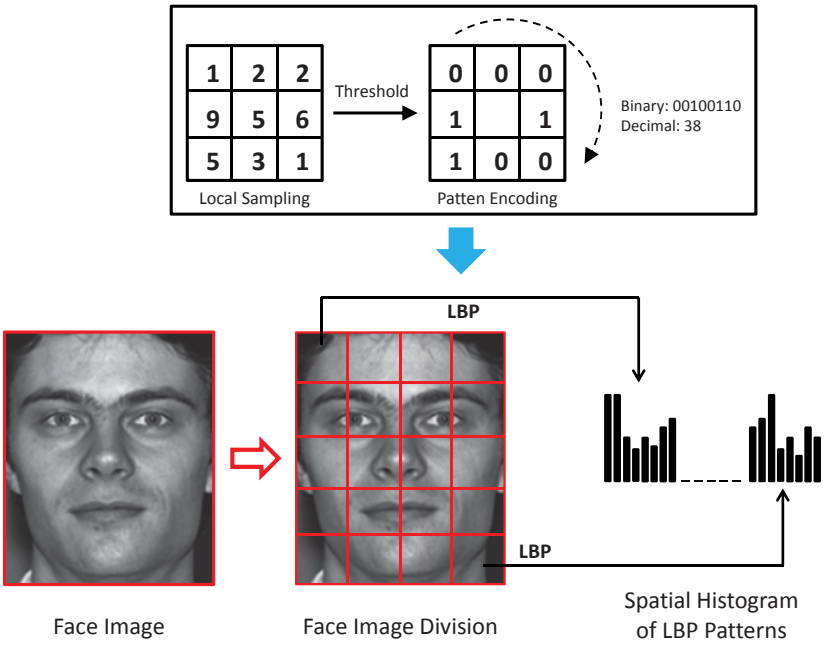


Figure 1.5: The principle of LBP-based face representation.

important [Ding et al., 2016].

1.2.2.1 Handcrafted Descriptors

Most local descriptors are hand-crafted and rely on expert knowledge. Of existing handcrafted descriptors, Local Binary Patterns (LBP) [Ahonen et al., 2004] and Gabor wavelets are two representative methods. The principle of LBP-based face representation is illustrated in Fig. 1.5. LBP encodes the gray-value differences between each pixel and its neighboring pixels into binary codes. A face image is then represented as the concatenated spatial histograms of the binary codes. Since encoding is determined by each pixel and its neighboring pixels, LBP encodes very detailed facial textures. Furthermore, the binarization operation helps to promote robustness to illumination variations.

LBP is extremely effective and efficient for face recognition. To further improve its performance, many LBP variants have been proposed including Local Ternary Patterns (LTP) [Tan and Triggs, 2010], Patterns of Oriented Edge Magnitudes (POEM) [Vu and Caplier, 2012], Transition LBP (tLBP) [Trefnỳ and Matas, 2010], Multi-Block

LBP (MB-LBP) [Zhang et al., 2007b], and LBP from Three Orthogonal Planes (LBP-TOP) [Zhao and Pietikainen, 2007]. In brief, LTP extracts ternary patterns that are more robust to face image noise. POEM extracts LBP patterns from oriented gradient maps instead of the original grayscale image, which contribute to illumination-robust face recognition. tLBP extracts complementary information to LBP by conducting neighboring pixel comparisons in a circular direction except for the central one. MB-LBP extracts multi-scale LBP patterns from local patches of different sizes. LBP-TOP captures dynamic grayscale variations from successive video frames.

Gabor wavelets aim to encode multi-scale and multi-orientation face image information. Therefore, compared to LBP and its variants, Gabor-based descriptors can extract information at larger scales. Representative Gabor-based descriptors include Local Gabor Binary Patterns (LGBP) [Zhang et al., 2005] and Local Gabor XOR Patterns (LGXP) [Xie et al., 2010]. Both methods adopt Gabor wavelets as a pre-processing step. LGBP encodes LBP patterns on Gabor magnitude maps, and because Gabor phase information is sensitive to translation, LGBP directly neglects phase information. To solve this problem, Xie et al. [2010] proposed extracting translation-robust XOR patterns using the phase maps and proved that Gabor magnitude maps and phase maps contain complementary information that, when combined, promotes face recognition performance. Gabor wavelets are gradient-like filters, so Gabor-based descriptors are usually robust to illumination variations. However, their efficiency is often poor due to the expensive Gabor convolution operations.

Another important handcrafted descriptor is Local Phase Quantization (LPQ) [Ahonen et al., 2008]. In contrast to the above descriptors, LPQ extracts blur-robust binary patterns from each local face image patch. Therefore, it has been widely employed for blur-robust face recognition, such as in video surveillance [Beveridge et al., 2015].

1.2.2.2 Learning-based Descriptors

Manually designing a good local descriptor requires substantial expert knowledge and experimental verification; therefore, the evolution of handcrafted descriptors has been slow. To solve this problem, learning-based descriptors have been proposed that make use of various machine learning models to automatically learn encoders from data. These methods have the following two advantages: first, they benefit from an

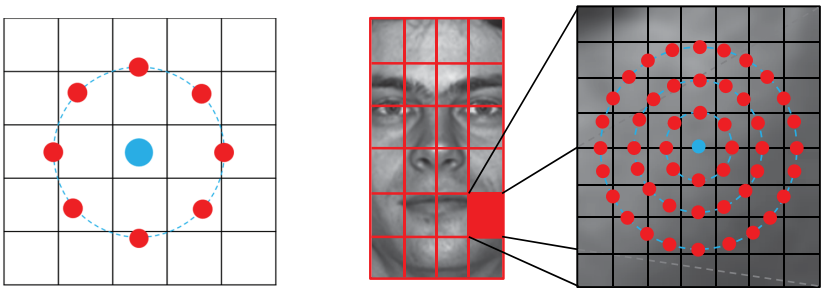


Figure 1.6: A comparison between handcrafted descriptors and learning-based descriptors during local sampling. (a) Handcrafted descriptors sample very limited numbers of pixels in the local patch; (b) learning-based descriptors can sample as many pixels as desired.

increasing volume of training data and progress in machine learning; and second, their encoders are flexible, i.e., they can process as many sampled points as desired, as illustrated in Fig. 1.6. In comparison, encoders in handcrafted descriptors usually process only a very limited number of sampled points. Therefore, learning-based descriptors can encode richer information. Representative learning-based descriptors include LE [Cao et al., 2010], Binarised Statistical Image Features (BSIF) [Kannala and Rahtu, 2012], Local Quantization Patterns (LQP) [Ul Hussain and Triggs, 2012], Discriminant Face Descriptor (DFD) [Lei et al., 2014], and PCANet [Chan et al., 2014].

Cao et al. [2010] proposed the LE descriptor, which employs the random projection tree (RPT) to learn encoders. Compared to traditional handcrafted descriptors, LE aims to produce uniformly distributed codes by making use of the RPT property. Similarly, Ul Hussain and Triggs [2012] adopted the K-means algorithm to learn encoders. Both RPT and K-means are unsupervised algorithms; therefore, the extracted features may not be sufficiently discriminative. To utilize training data label information, Lei et al. [2014] proposed the DFD descriptor, which employs LDA to learn encoders.

Local sampling as illustrated in Fig. 1.6 may cause information loss. Therefore, BSIF [Kannala and Rahtu, 2012] skips the local sampling step and directly employs all the pixels within local patches. A set of gradient-like filters learnt by Independent Component Analysis (ICA) and a simple binarization operation form the encoder.

Similarly, PCANet [Chan et al., 2014] learns the filters by PCA. Their main difference is that PCANet is deeper than BSIF: PCANet learns two sets of PCA filters and applies them successively to a face image; therefore, PCANet can encode more abstract texture information than BSIF.

1.2.3 Deep Learning-based Representations

Both subspace learning-based methods and local descriptor-based methods have shallow structures; therefore, they do not represent highly abstract and nonlinear information very well. In comparison, deep learning-based approaches have the advantages of hierarchy and a deep structure. The most successful deep learning models for face recognition are those that use Convolutional Neural Networks (CNNs).

Early CNN approaches for face recognition date back to the last century [Lawrence et al., 1997]. However, due to a large number of model parameters and an absence of large volumes of training data, early CNN-based approaches were difficult to optimize, and the community abandoned CNN-based approaches for over a decade. Fortunately, with the arrival of big data and improvements in CNN model structure, CNNs can now be effectively optimized to learn representations that are both highly discriminative and generalizable. In fact, CNN-based methods have dominated the face recognition field over the last two years. We briefly review some representative CNN-based approaches for facial feature extraction.

1.2.3.1 CNN Model Architectures

Representative CNN models proposed by the face recognition community include Deep Face [Taigman et al., 2014a] and DeepID [Sun et al., 2014], et al. Recently, the face recognition community has benefited from new CNN architectures for image classification. The top performers in image classification, e.g., the VGG model [Simonyan and Zisserman, 2014] and GoogLeNet [Szegedy et al., 2015], have been modified and deployed for face recognition with excellent results [Parkhi et al., Schroff et al., 2015].

1.2.3.2 Deep Metric Learning

Typically, face representation learning in CNN is formulated as a face identification problem using the softmax loss function. However, the softmax loss function targets the classification error of training images rather than the quality of learnt face representations. To solve this problem, deep metric learning models have been employed to optimize CNN parameters and promote the discriminative power of learnt face representations. Popular deep metric learning models include pairwise loss [Sun et al., 2014] and triplet loss [Schroff et al., 2015], with improved performance observed in [Ding and Tao, 2015, Sun et al., 2014].

1.2.3.3 Prospects

The high-quality face representations learnt by CNN models have resulted in nearly perfect performance on some traditionally challenging databases such as the Labeled Faces in the Wild (LFW) database [Ding and Tao, 2015, Liu et al., 2015, Schroff et al., 2015]. However, this does not mean that face recognition has been completely solved. Instead, researchers have applied CNN-based approaches to solve even more difficult face recognition problems such as large-scale face recognition [Kemelmacher-Shlizerman et al., 2015], pose-invariant face recognition [Ding and Tao, 2016], and video-based face recognition [Beveridge et al., 2015].

1.3 Classification Models

Face classification is based on the similarity score between face representations, and the similarity score is computed by a certain distance metric. Simple handcrafted metrics exist including the Euclidean distance, cosine distance, and chi-square distance, while more discriminative metrics are usually learnt from labeled data that can be divided into two categories: discriminative models and generative models.

1.3.1 Discriminative Models

The majority of discriminative models fall within the scope of Mahalanobis metrics. Popular methods include LDA, Large Margin Nearest Neighbor (LMNN) [Weinberger

et al., 2005], Information-theoretic Metric Learning (ITML) [Davis et al., 2007], Logistic Discriminant-based Metric Learning (LDML) [Guillaumin et al., 2009], and Pairwise Constrained Component Analysis (PCCA) [Mignon and Jurie, 2012].

Different models differ in the detail of their loss functions and constraints. However, their basic principle can be formulated as follows. Given two face representations $x_i \in \mathbb{R}^{D\times 1}$ and $x_j \in \mathbb{R}^{D\times 1}$, the object of one Mahalanobis metric is to learn a symmetric positive definite matrix $M \in \mathbb{R}^{D\times D}$:

$$d_M(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j), \tag{1.2}$$

such that $d_M(x_i, x_j)$ of two intra-class face presentations is small while $d_M(x_i, x_j)$ of two inter-class face presentations is large.

1.3.2 Generative Models

Of the generative models, Wright et al. [Wright et al., 2009] proposed the Sparse Representation-based Classification (SRC) approach for face classification. Based on compressed sensing theory, SRC represents each test face as a sparse and linear combination of gallery (training) samples. If the test (probe) face belongs to one subject in the training (gallery) set, then the coefficients of the linear combination should concentrate on samples of the same subject as the test face; the coefficients of the samples belonging to different subjects should be close to zero. The similarity score is obtained by calculating the reconstruction error between the test face and the linear combination of samples from each subject. SRC has shown great success on some simple face databases, but a major limitation of SRC-based approaches is that they require multiple samples for each subject in the training (gallery) set. However, many real-world circumstances only contain one sample for each training (gallery) subject, as argued in [Tan et al., 2006].

Other generative models include probabilistic approaches, e.g., Probabilistic LDA (PLDA) [Prince and Elder, 2007] and Joint Bayesian (JB) [Chen et al., 2012a]. PLDA models the face data generation process as:

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \varepsilon_{ij}, \tag{1.3}$$

where x_{ij} denotes the jth face data of the ith individual. μ is the mean of all face data. F and G are low-rank matrices whose columns are the basis vectors of the between-individual subspace and the within-individual subspace, respectively. h_i is the latent identity variable that is constant for all images of the ith subject. w_{ij} and ε_{ij} are noise terms explaining intra-personal variance. It is assumed that h_i, w_{ij}, and ε_{ij} follow Gaussian distributions. The Expectation-Maximization (EM) algorithm estimates the above model parameters from labeled face data.

In practice, additional user involvement is required to tune the dimensions of F and G, because their intrinsic dimensionality is usually unknown. In contrast, JB [Chen et al., 2012a] frees itself from this hyper-parameter by simplifying the face data generation process:

$$x_{ij} = \mu_i + \varepsilon_{ij}, \tag{1.4}$$

where μ_i represents the identity of the ith subject, and ε_{ij} represents intra-personal noise. Both μ_i and ε_{ij} follow Gaussian distributions. Based on the simplified face prior, JB models the joint distribution of two face images $\{x_1, x_2\}$. Similar to PLDA, JB's model parameters are estimated from labeled face data using the EM algorithm. During testing, the log-likelihood ratio of whether two face images belong to the same subject or not can be utilized as similarity scores for face recognition in both PLDA and JB.

1.4 Pose-invariant Face Recognition

The appearance changes caused by pose variation often significantly surpass intrinsic differences between individuals. As a consequence, it is often not possible or effective to directly compare two images under different poses as in conventional face recognition algorithms. Explicit strategies are required to bridge the cross-pose gap; therefore, pose-invariant face recognition (PIFR) approaches have developed as an independent research topic.

PIFR refers to the problem of identifying or authorizing individuals with face images captured under arbitrary poses, as shown in Fig. 1.7. The first explorations of PIFR date back to the early 1990s [Beymer, 1994, Brunelli and Poggio, 1993, Pentland et al., 1994]. Nevertheless, the substantial facial appearance changes caused by pose

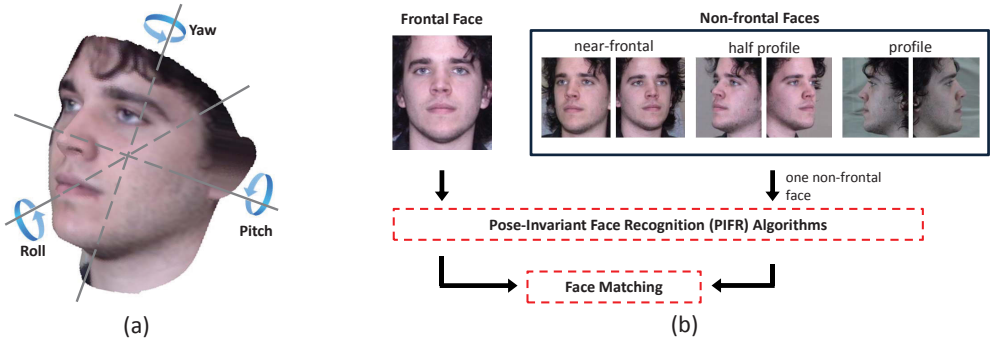


Figure 1.7: (a) The three degrees of freedom of face pose variation: yaw, pitch, and roll. (b) A typical PIFR framework. Different to the traditional near-frontal face recognition (NFFR), PIFR aims to recognize faces captured under arbitrary poses.

variation continue to challenge even state-of-the-art face recognition systems. This is mainly a result of the complex 3D structure of the human head, which presents the following specific challenges (see also Fig. 1.8):

- The rigid rotation of the head results in self-occlusion, which means there is loss of information for recognition.
- The position of facial textures varies nonlinearly following pose changes, which represents a loss of semantic correspondence in 2D images.
- The shape of facial textures warps nonlinearly along with pose changes, which causes serious confusion with inter-personal texture differences.
- The pose variation is usually combined with other factors to simultaneously affect face appearance. For example, far-away subjects tend to exhibit larger pose variations as they are unaware of the cameras. Therefore, low resolution and illumination variations occur together with large pose variations.

A wide variety of approaches have been proposed over recent years that can broadly be grouped into four categories that handle PIFR from distinct perspectives:

- Those that extract pose-robust features as face representations, allowing conventional classifiers to be employed for face matching.

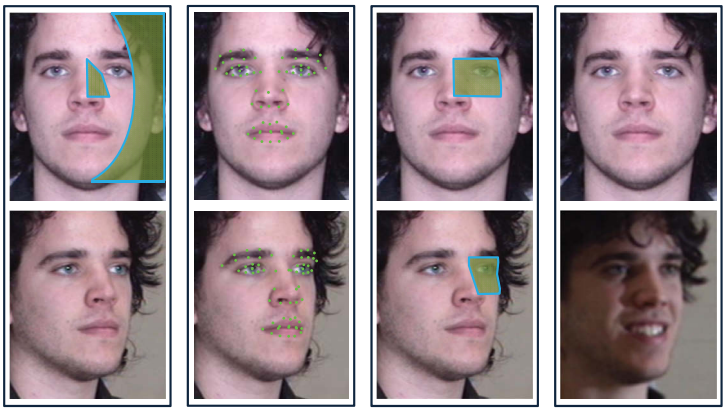


Figure 1.8: The challenges for face recognition caused by pose variation. (a) Self-occlusion: the marked area in the frontal face is invisible in the non-frontal face; (b) loss of semantic correspondence: the position of facial textures varies nonlinearly following the pose change; (c) nonlinear warping of facial textures; (d) accompanying variations in resolution, illumination, and expression.

- Those that project features of different poses into a shared latent subspace where face matching is meaningful.
- Those that synthesize face images from one pose to another pose, so that the two faces originally in different poses can be matched in the same pose with traditional frontal face recognition algorithms.
- Those that combine two or three of the above techniques for more effective PIFR.

Inspired by [Ouyang et al., 2014], we unify the four PIFR approaches in the following formulation:

$$M\left[W^a F\left(S^a(\mathbf{I}_i^a)\right), W^b F\left(S^b(\mathbf{I}_j^b)\right)\right], \tag{1.5}$$

where $\mathbf{I}_i^a$ and $\mathbf{I}_j^b$ stand for two face images in pose a and pose b, respectively; S^a and S^b are synthesis operations, after which the two face images are under the same pose; F denotes pose-robust feature extraction; W^a and W^b correspond to feature transformations learnt by multi-view subspace learning algorithms; and M means a face matching algorithm, e.g., the nearest neighbor (NN) classifier. It is easily appreciated that the first three approaches focus their effort on only one operation in

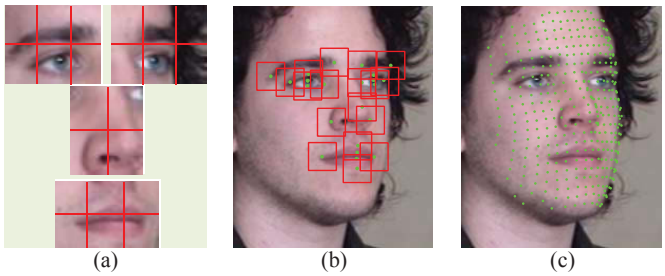


Figure 1.9: Feature extraction from semantically corresponding patches or landmarks. (a) Semantic correspondence realized at the facial component level [Brunelli and Poggio, 1993, Pentland et al., 1994]; (b) semantic correspondence by detecting dense facial landmarks [Chen et al., 2013, Ding et al., 2016, Wiskott et al., 1997]; (c) tight semantic correspondence realized using various techniques, e.g., 3D face model [Li et al., 2009, Yi et al., 2013] and MRF [Arashloo and Kittler, 2011].

Eq. 1.5. For example, the multi-view subspace learning approaches provide strategies for determining the mappings W^a and W^b, while the face synthesis-based methods are devoted to solving S^a and S^b. The hybrid approaches may contribute to two or more steps in Eq. 1.5. The four approach categories are briefly reviewed below.

1.4.1 Pose-robust Feature Extraction

Approaches in this category focus on designing face representations intrinsically robust to pose variation while remaining discriminative to the identity of subjects. Depending on whether the features are extracted by manually designed descriptors or by trained machine learning models, the approaches in this category can be grouped into engineered features and learning-based features.

The engineered features handle pose challenges by explicitly re-establishing the semantic correspondence during feature extraction, as illustrated in Fig 1.9. Early PIFR approaches [Brunelli and Poggio, 1993, Pentland et al., 1994] realized semantic correspondence across poses at the facial component level by detecting sparse facial component landmarks, e.g., both eye centers, the nose tip, and the mouth center. The set of facial component-level features compose the pose-robust face representation. More recent engineered features benefit from the rapid progress made in facial landmark detection, which makes dense landmark detection more reliable. For example, Chen et al. [2013] extracted multi-scale Local Binary Patterns (LBP) features

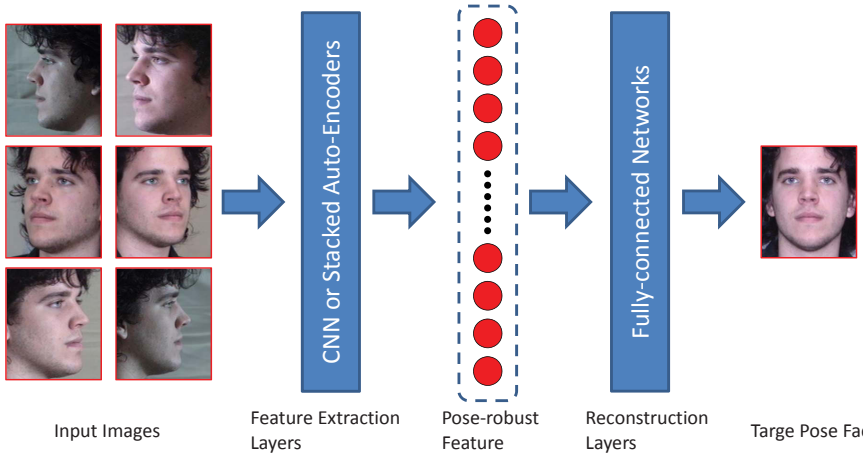


Figure 1.10: The common framework of deep neural network-based pose-robust feature extraction methods [Kan et al., 2014, Zhang et al., 2013, Zhu et al., 2013].

from patches around 27 landmarks, the LBP features for all patches then being concatenated to produce a high-dimensional feature vector as the pose-robust feature.

Some approaches are landmark detection free. For example, Arashloo et al. [2011] proposed an approach based on Markov Random Field (MRF) to match semantically corresponding patches between two images. In this approach, the densely sampled image patches are represented as nodes in the MRF model, while the 2D displacement vectors are treated as labels. The goal of MRF-based optimization is to assign labels at minimum cost while taking both translations and projective distortions into consideration.

The learning-based features learn pose-robust features from multi-pose training data. For example, Zhu et al. [2013] employed deep neural networks to extract pose-robust features, as illustrated in Fig. 1.10. The deep network is the stack of two main modules: the feature extraction module and the frontal face reconstruction module. The model input is a set of pose-varied images of an individual, and the output of the feature extraction module is employed to recover frontal faces through the latter module. The logic of this method is that, regardless of the pose of the input image, the output of the reconstruction module is encouraged to be as close as possible to the frontal pose image of the subject. Thus, the output of the feature extraction module must be pose-robust. Due to the deep model structure, the network must tune millions of parameters and therefore requires a large amount of multi-pose training data.

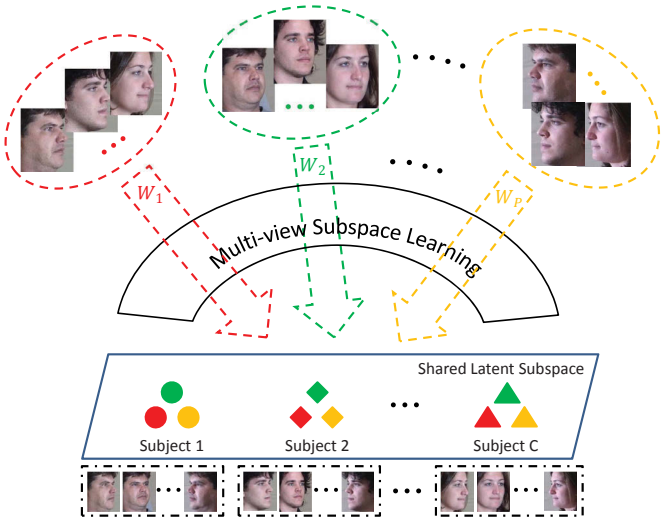


Figure 1.11: The framework of multi-view subspace learning-based PIFR approaches [Kan et al., 2012, Li et al., 2009, Prince et al., 2008, Sharma et al., 2012]. The continuous pose range is divided into P discrete pose spaces, and pose-specific projections (i.e., $W_1, W_2, \cdots, W_P$) to the latent subspace are learnt.

1.4.2 Multi-view Subspace Learning

Pose-varied face images are distributed on a highly nonlinear manifold [Tenenbaum et al., 2000], which greatly degrades the performance of traditional face recognition models based on the single linear subspace assumption. The idea of multi-view subspace learning-based approaches dates back to [Kim et al., 2003, Prince and Elder, 2005], who proposed dividing the nonlinear manifold into a discrete set of pose spaces and treated each pose as a single view; pose-specific projections to a latent subspace shared by different poses are subsequently learnt. Since the images of one subject are captured under different poses of the same face, they should be highly correlated in this subspace; therefore, face matching can be performed due to feature correspondence. An illustration of the multi-view subspace learning framework is shown in Fig 1.11.

Representative models include Canonical Correlation Analysis (CCA) [Li et al., 2009], Partial Least Squares (PLS) [Sharma and Jacobs, 2011], Generalized Multiview Analysis (GMA) [Sharma et al., 2012], and their nonlinear extensions [Akaho, 2006]. However, these methods usually require the same number of faces for all poses of each training subject, a condition that may not be satisfied in real applications. To handle this problem, Kan et al. [2012] proposed the Multi-view Discriminant Analysis

(MvDA) approach which utilizes all face images from all poses. MvDA builds a single between-class scatter matrix S_b and a single within-class scatter matrix S_w from both the inter-pose and intra-pose embeddings in the shared subspace:

$$\begin{aligned} S_w &= \sum_{i=1}^{C} \sum_{j=1}^{P} \sum_{k=1}^{n_{ij}} (y_{ij}^k - \mu_i)(y_{ij}^k - \mu_i)^T, \\ S_b &= \sum_{i=1}^{C} n_i (\mu_i - \mu)(\mu_i - \mu)^T, \end{aligned} \tag{1.6}$$

where y_{ij}^k stands for the low-dimensional embedding of the k^{th} sample from the j^{th} pose of the i^{th} subject. μ_i is the mean of the low-dimensional embeddings of the i^{th} subject, and μ is the mean of the low-dimensional embeddings of all C subjects. The objective of MvDA is similar to that of LDA, i.e., to maximize the ratio between S_b and S_w. The pose-specific projections are obtained by formulating the objective as the optimization of a generalized Rayleigh quotient. As variations from both inter-pose and intra-pose faces are considered together in the objective function, the authors argued that a more discriminative subspace can be learnt.

1.4.3 Face Synthesis Based on 2D Methods

Since it is difficult to directly match two faces under different poses, an intuitive approach is to perform face synthesis such that the two faces are transformed to the same pose, allowing conventional face recognition algorithms to be used for matching. Existing PIFR face synthesis methods can be broadly classified into methods based on 2D techniques and methods based on 3D techniques, depending on whether the synthesis is accomplished in the 2D or 3D domain. In this subsection, we review the 2D techniques.

1.4.3.1 2D Pose Normalization

Three main 2D pose normalization schemes are illustrated in Fig. 1.12: piece-wise warping [Cootes et al., 2001], patch-wise warping [Ashraf et al., 2008], and pixel-wise displacement [Beymer and Poggio, 1995, Li et al., 2012c]. These methods synthesize faces by calculating the warps across poses for each piece, patch, or pixel in a face image. Their advantage is that they require only limited or no training data, and they

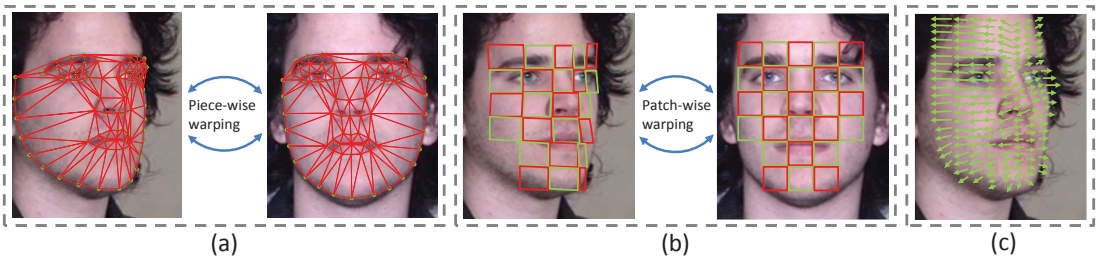


Figure 1.12: Three main 2D-based pose normalization schemes. (a) Piece-wise warping; (b) patch-wise warping; and (c) pixel-wise displacement.

preserve the fine textures of the original image. However, the downside is that they cannot recover occluded facial textures.

1.4.3.2 Linear Regression Models

Same as in the approaches described in Section 1.4.2, the methods in this subsection divide the continuous pose space into a set of discrete pose segments. Faces that fall into the same pose segment are assumed to have the same pose p.

The earliest work to formulate face synthesis as a linear regression problem was Beymer and Poggio [1995]. Under an assumption of orthogonal projection and constant illumination, the holistic face image is represented as the linear combination of a set of training faces in the same pose, and the same combination coefficients are employed for face synthesis under another pose. However, this approach requires dense pixel-wise correspondence between face images, which is challenging in practice. Later works conducted face synthesis using a patch-wise strategy [Chai et al., 2007], thereby reducing the difficulty in alignment.

1.4.3.3 Nonlinear Regression Models

Appearance variations of face images across poses are intrinsically nonlinear due to substantial occlusion and nonlinear warp. To synthesize higher-quality face images, nonlinear regression models have recently been introduced.

The works reviewed in Section 1.4.1 [Kan et al., 2014, Zhang et al., 2013, Zhu et al., 2013] adopt neural networks as nonlinear regression models for 2D face synthesis by virtue of their power to learn nonlinear transformations. These four works

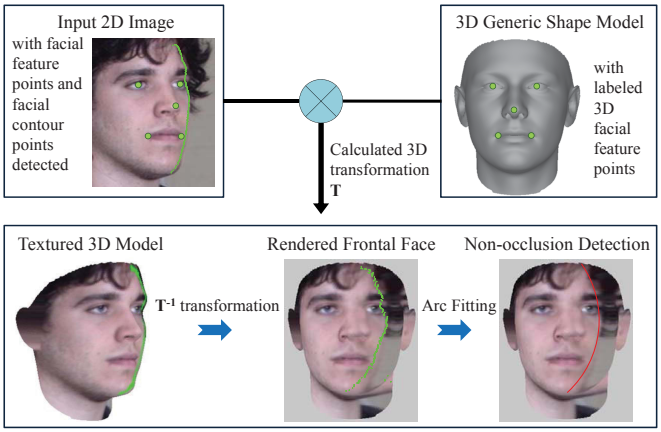


Figure 1.13: The pipeline for 3D pose normalization from a single face image proposed in [Ding et al., 2015]. Face regions that are free from occlusion are detected and employed for face recognition.

share the common strategy that pose-robust feature is first extracted and then utilized for frontal face recovery.

1.4.4 Face Synthesis Based on 3D Methods

The human head is a complex non-planar 3D structure rotating in 3D space, while a face image lies in the 2D domain. The lack of one degree of freedom makes accurate face synthesis using 2D techniques difficult. The methods reviewed in this subsection build a 3D model of the human head and then synthesize faces based on the 3D face models. The 3D methods can be classified into 3D pose normalization and 3D modeling by image reconstruction.

1.4.4.1 3D Pose Normalization

The 3D pose normalization approaches employ the 3D facial shape model as a tool to correct the nonlinear warping of facial textures appearing in 2D images. Like the 2D pose normalization methods reviewed in Section 1.4.3, these methods preserve the original pixel values of the input image. As illustrated in Fig. 1.13, the general principle is that the 2D face image is first aligned with a 3D face model, typically with the help of facial landmarks [Hassner, 2013, Jiang et al., 2005]. Then, the texture of the 2D image is mapped to the 3D model. Finally, the textured 3D model is rotated

to a desired pose and a new 2D image in that pose is rendered. Early approaches utilized simple 3D models, e.g., the cylinder model [Gao et al., 2001], the wire frame model [Lee and Ranganath, 2003, Zhang et al., 2006], and the ellipsoid model [Liu and Chen, 2005], to roughly model the 3D structure of the human head, whereas newer approaches strive to build accurate 3D facial shape models.

1.4.4.2 3D Modeling by Image Reconstruction

The 3D pose normalization approaches take their cues from a set of facial landmarks for 3D shape reconstruction. The information contained in the facial landmarks is limited, making accurate 3D face reconstruction and subsequent face synthesis difficult. In contrast, the approaches reviewed in this subsection make full use of every pixel in the image to infer the 3D structure of the face image.

The most classical approach is the 3D Morphable Model (3DMM)[Blanz and Vetter, 2003]. 3DMM simulates image formation by combining a deformable 3D model and computer graphics techniques. Given a face image, 3DMM automatically estimates the 3D shape coefficient vector α, texture coefficient vector β, and parameters of computer graphics models γ by fitting the deformable 3D model with the face image. The model parameters are optimized using a stochastic version of Newton's algorithm, the objective being that the sum of squared differences over all pixels between the rendered image and the input image should be as similar as possible.

Ideally, the separated shape and texture parameters α and β are only related to identity and thus provide pose and illumination invariance. Face recognition can therefore be conducted by comparing α and β between gallery and probe images. Also, with the optimized α and β, 2D face images can be synthesized under arbitrary poses for the subject appearing in the input image. The major disadvantage of 3DMM lies in its fitting procedure, which is highly nonlinear and prone to trapping in local minima; therefore, the synthesized face image may not be accurate.

1.4.5 Hybrid Methods

Hybrid methods combine one or more of the aforementioned PIFR strategies to make use of the complementary advantages of the different methods. The hybrid approaches are less well studied in the literature but tend to be more powerful than any single

PIFR category alone. Therefore, they hold more promise for solving real-world PIFR problems. Several successful combinations are reviewed below.

A number of approaches combine pose-robust feature extraction and multi-view subspace learning [Fischer et al., 2012, Prince et al., 2008]. Instead of extracting holistic features from the whole face image, they extract pose-robust features around facial landmarks, substantially reducing the difficulty of multi-view subspace learning. This strategy has been shown to significantly enhance the performance of PIFR systems. Ding et al. [2015] proposed a combination of 3D-based face synthesis and multi-view subspace learning inspired by the fact that frontal faces synthesized from non-frontal images of distinct poses differ in image quality, an aspect that needs to be improved by multi-view subspace learning. It is also possible to use two or more categories of techniques independently and fuse their estimates into a single result. For example, Kim and Kittler [2006] proposed an expert fusion system in which the pose-robust feature extraction expert, multi-view subspace learning expert, and face synthesis expert are run independently. Expert results are then fused at the score level to produce impressive performance improvements.

1.4.6 Relationships Between the Four Categories

In this subsection, we discuss the relative strengths and limitations of the different categories.

High quality face representation is critical for the traditional NFFR problem, and deep learning-based methods have been highly successful for representation learning [Schroff et al., 2015, Taigman et al., 2015]. In the case of PIFR, we expect the pose-robust features extracted by powerful deep models to continue to play a critical role, provided that massive amounts of labeled multi-pose training data exist. However, multi-pose training data large enough to drive complicated deep models may be difficult to collect in real-world applications.

It is not always easy to independently exploit multi-view subspace learning methods, because they are based on the ideal assumption that simple pose-specific projections eliminate cross-pose face differences. Furthermore, their performance is highly dependent on the amount of labeled multi-pose training data. In practice, multi-view subspace learning methods should be combined with pose-robust features

to reduce the cross-pose gap.

Pose normalization-based face synthesis strategies are particularly useful when there is no multi-pose training data or the training data are small. Their main limitation is artifacts created in the synthesized image by the inaccurate estimation of facial shape or pose parameters. These artifacts change the original appearance of the subject and thus deteriorate the subsequently extracted features, causing an adverse impact on high-precision face recognition, as empirically found in a recent work [Ding and Tao, 2015].

Therefore, both pose-robust feature extraction and pose normalization are promising solutions to PIFR. In practice, the choice of the most appropriate PIFR method mainly depends on the availability and size of multi-pose training data. The degree of pose variation is another important factor to consider: for near-frontal or half-profile face images, existing pose-robust feature extraction methods are highly accurate; for profile face images, the synthesis-based approaches may be more useful for bridging the huge appearance gap between poses. Moreover, hybrid methods try to solve the PIFR problem from multiple perspectives; therefore, they may be more promising for real-world PIFR problems.

1.5 Face Databases

The development of face recognition technology is closely related to the availability of face databases. On the one hand, public face databases provide a platform for the fair comparison of different face recognition algorithms. On the other hand, they reveal the shortfalls in existing algorithms through failed cases and thus provide the rationale to develop new algorithms. With the performance of new face recognition algorithms saturating on old databases, more difficult face databases are published to advance the field, as illustrated in Fig. 1.14. For example, in the early stages of face recognition, face databases tended to be small and their images were captured under controlled (laboratory) conditions. Nowadays, face recognition in the controlled environment has largely been solved, and images in new face databases are all captured in uncontrolled (real-world) conditions. Some representative face databases are briefly introduced in the following sections. They can be classified into two categories: still face image databases and video face databases.

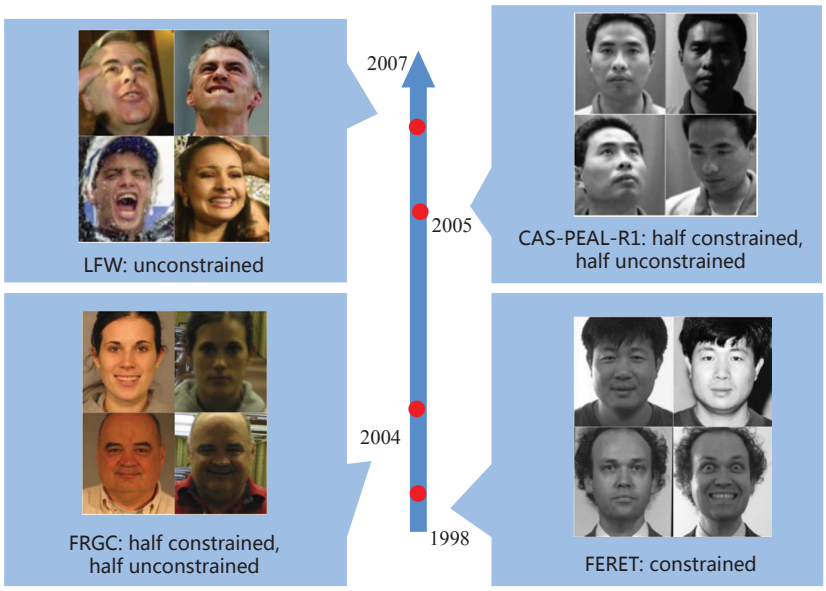


Figure 1.14: The evolution of face databases.

1.5.1 Still Face Image Databases

FERET: The FERET [Phillips et al., 2000] database was designed to evaluate the performance of face identification algorithms under controlled conditions. FERET frontal images are divided into one gallery set Fa (1,196 images of 1,196 subjects) and four probe sets: Fb (1,195 images with different facial expressions), Fc (194 images captured under different illumination conditions), Dup1 (722 images captured a week later), and Dup2 (234 images captured a year later).

FRGC 2.0: The FRGC 2.0 [Phillips et al., 2005] database was constructed to evaluate the performance of face verification algorithms in both controlled and less controlled environments. Images in FRGC 2.0 are partitioned into a training set and a test set. The training set is composed of 12,776 images from 222 subjects, with one half controlled still images and the other half uncontrolled still images. The test set is made up of 24,042 images of 466 subjects - 16,028 controlled still images and 8,014 uncontrolled still images. Six experiments are defined in [Phillips et al., 2005], but usually only Experiments 1 and 4 are tested. The shared target set of Experiments 1 and 4 are collected from frontal facial images taken under controlled illumination, while images in their query sets are captured under controlled and less controlled conditions,

respectively.

CAS-PEAL-R1: The CAS-PEAL-R1 [Gao et al., 2008] database allows researchers to test the robustness of their algorithms to particular sources of variation. Images in CAS-PEAL-R1 are divided into one training set (1,200 frontal images of 300 subjects), one gallery set (1,040 images of the 1,040 subjects), and nine probe sets. Each of the nine probe sets is restricted to one main variation. In detail, the first six probe sets PE, PA, PL, PT, PB, and PS correspond to variations in expression, accessory, lighting, time, background, and distance of frontal face images, respectively. The remaining three probe sets PU, PM, and PD correspond to one type of pose variation.

Multi-PIE: The Multi-PIE database [Gross et al., 2010] contains images of 337 subjects from 15 different viewpoints, 19 illumination conditions, and up to 6 expression types. All images were captured under laboratory conditions. This database is usually utilized to evaluate the robustness of face recognition algorithms to a certain influential factor, e.g., pose or illumination variations.

LFW: The LFW [Huang et al., 2007] database was created for unconstrained face verification research. It contains 13,233 images of 5,749 subjects. In contrast to the previous databases, whose images were captured under laboratory conditions, all images in LFW were collected from internet webpages; therefore, the LFW images feature dramatic intra-personal variations in occlusion, illumination, and expression. However, faces in LFW were detected automatically by the simple Viola-Jones face detector [Viola and Jones, 2004], which constrains the face pose range in LFW.

IJB-A: The IARPA Janus Benchmark A (IJB-A) database [Klare et al., 2015] is a newly published face database containing 5,712 face images and 2,085 videos from 500 subjects. Similar to LFW, images in IJB-A database are collected from the internet. The key characteristic of IJB-A is that both face detection and facial feature point detection are accomplished manually. Therefore, face images in IJB-A database cover the full range of pose variations.

1.5.2 Video Face Databases

YouTube Faces: The YouTube Faces database [Wolf et al., 2011a] includes 3,425 videos of 1,595 subjects. These subjects are a subset of the LFW database. All

videos were downloaded from the YouTube website. As the majority of subjects in this database were in interviews, the intra-personal appearance variance within each video is not significant.

PaSC: The PaSC database [Beveridge et al., 2013] contains 2,802 videos of 265 subjects. The videos were recorded by multiple sensors in varied indoor and outdoor locations and are divided into two sets: a control set and a handheld set. Videos in the control set were captured by a high-end camera installed on a tripod, and their image quality is relatively good. Videos in the handheld set were captured by five handheld video cameras. For each video, the subject was asked to carry out actions in order to present a wide range of poses at variable distances to the camera. Faces in the video exhibit serious motion and out-of-focus blur and rich pose variations.

COX Face: The COX Face database [Huang et al., 2015b] incorporates 1,000 still images and 3,000 videos of 1,000 subjects. The still images were captured by a high quality camera under controlled conditions to simulate ID photos. The videos were taken while the subjects were walking in a large gym to simulate a surveillance environment. Three cameras from different locations were installed to capture videos of the walking subject simultaneously.

1.6 Contributions and Related Publications

This thesis studies **Robust Face Recognition**. We: (i) investigate the impact of major challenges in face recognition, e.g., pose variation, illumination variation, expression variation, image blur, and occlusion; and (ii) propose a series of novel face recognition algorithms that are robust to these factors by taking advantage of advanced machine learning techniques including multi-task learning, metric learning, and deep learning.

Specifically, in Chapter 2 & Chapter 3, we first study new face image descriptors for generic face recognition purposes, since face image descriptors are the core technology for effective face representation. In Chapter 4 & Chapter 5, based on our research on face image descriptors, we further study the two most challenging applications in face recognition: pose-invariant face recognition (Chapter 4) and video-based face recognition (Chapter 5). The proposed algorithms are superior to existing algorithms. The organization of this thesis is illustrated in Fig. 1.15, and the detailed contribution of each chapter is as follows.

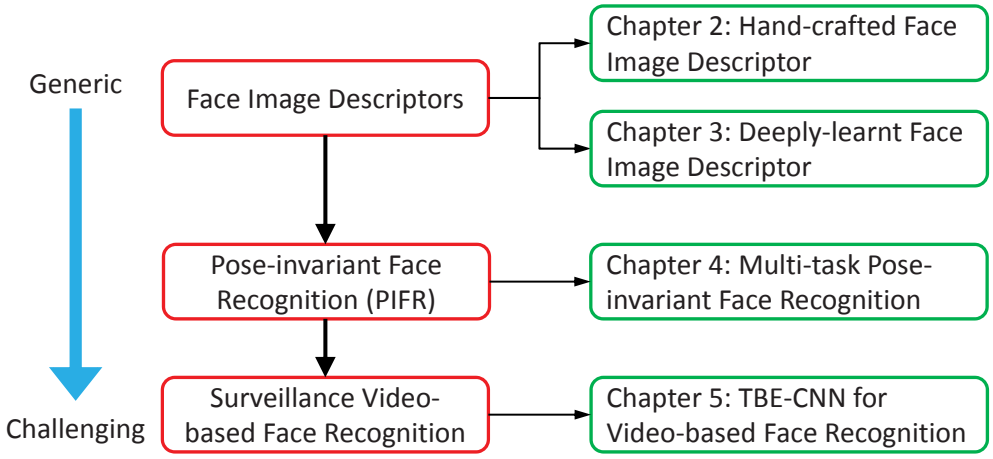


Figure 1.15: Structure of this thesis.

In Chapter 2, we manually design a novel face image descriptor named "Dual-Cross Patterns" (DCP). DCP encodes the second-order statistics of facial textures in the most informative directions within a face image; therefore, DCP is more descriptive and discriminative than existing handcrafted face image descriptors. We also make DCP efficient by employing a dual-cross grouping strategy based on the maximum joint Shannon entropy principle; therefore the time cost of DCP is only twice that of LBP. We further extend DCP into a comprehensive face representation scheme named "Multi-Directional Multi-Level Dual-Cross Patterns" (MDML-DCPs). MDML-DCPs efficiently encodes the invariant characteristics of a face image from multiple levels into patterns that are highly discriminative of inter-personal differences but robust to intra-personal variations. MDML-DCPs achieves the best performance on four large-scale face databases: FERET, FRGC 2.0, CAS-PEAL-R1, and LFW.

In Chapter 3, we develop a deep learning-based face image descriptor named "Multimodal Deep Face Representation" (MM-DFR) to automatically learn face representations from multimodal image data. In MM-DFR, a set of CNNs are designed to extract multimodal information from the original holistic face image, the frontal pose image rendered by 3D modeling, and uniformly sampled image patches. A feature-level fusion approach using stacked auto-encoders is designed to fuse the features extracted from the set of CNNs, which is advantageous for nonlinear dimension reduction. MM-DFR achieves greater than 99% recognition rate on LFW using publicly available training set.

In Chapter 4, based on our research on handcrafted face image descriptors, we handle the challenging PIFR problem. We propose a powerful PIFR framework capable of handling the full range of pose variations within $\pm 90^{\circ}$ of yaw. Under the proposed framework, we elegantly transform the original PIFR problem into a partial frontal face recognition problem. DCP descriptors are extracted from non-occluded patches to represent the partial frontal face. We also propose a novel multi-task learning-based approach for PIFR to promote the discriminative power of face representation. The complete framework outperforms state-of-the-art PIFR methods on the FERET, CMU-PIE, and Multi-PIE databases.

In Chapter 5, based on our research on deep learning-based descriptors, we handle the surveillance video-based face recognition problem, which may be the most challenging application of face recognition. We focus on three major challenges: image blur, occlusion, and pose variations. First, to learn blur-robust face representations, we introduce artificial blur to the training data composed of clear still images, which effectively overcomes the shortfall in real-world video training data and encourages CNN to learn blur-insensitive features automatically. Second, to enhance the robustness of CNN features to pose variations and occlusion, we propose the TBE-CNN model, which extracts complementary information from the holistic face image and patches cropped from around facial components. We also make TBE-CNN efficient by sharing low- and middle-level convolutional layers of multiple networks. Third, we propose a new deep metric learning model to enhance the discriminative power of TBE-CNN. With the proposed framework, state-of-the-art performance is achieved on three popular video face databases: PaSC, COX Face, and YouTube Faces.

Publications related to this thesis are as follows.

- **Changxing Ding** and Dacheng Tao: A comprehensive survey on pose-invariant face recognition, ACM Transactions on Intelligent Systems and Technology (**TIST**), 7(3):1-42, 2016.

- **Changxing Ding**, Jonghyun Choi, Dacheng Tao, and Larry S Davis: Multi-directional multi-level dual-cross patterns for robust face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 38(3):518-531, 2016.

- **Changxing Ding** and Dacheng Tao: Robust face recognition via multimodal

deep face representation, IEEE Transactions on Multimedia (**TMM**), 17(11):2049-2058, 2015.

- **Changxing Ding**, Chang Xu, and Dacheng Tao: Multi-task pose-invariant face recognition, IEEE Transactions on Image Processing (**TIP**), 24(3):980-993, 2015.

- **Changxing Ding** and Dacheng Tao: Trunk-Branch Ensemble Convolutional Neural Networks for Video-based Face Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**, under review), 2016.

Chapter 2

Multi-Directional Multi-Level Dual-Cross Patterns

To perform unconstrained face recognition robust to variations in illumination, pose and expression, this chapter presents a new scheme to extract "Multi-Directional Multi-Level Dual-Cross Patterns" (MDML-DCPs) from face images. Specifically, the MDML-DCPs scheme exploits the first derivative of Gaussian operator to reduce the impact of differences in illumination and then computes the DCP feature at both the holistic and component levels. DCP is a novel face image descriptor inspired by the unique textural structure of human faces. It is computationally efficient and only doubles the cost of computing local binary patterns, yet is extremely robust to pose and expression variations. MDML-DCPs comprehensively yet efficiently encodes the invariant characteristics of a face image from multiple levels into patterns that are highly discriminative of inter-personal differences but robust to intra-personal variations.

Experimental results on the FERET, CAS-PERL-R1, FRGC 2.0, and LFW databases indicate that DCP outperforms the state-of-the-art local descriptors (e.g. LBP, LTP, LPQ, POEM, tLBP, and LGXP) for both face identification and face verification tasks. More impressively, the best performance is achieved on the challenging LFW and FRGC 2.0 databases by deploying MDML-DCPs in a simple recognition scheme.

2.1 Introduction

Face recognition has been an active area of research due to both the scientific challenge and its potential use in a wide range of practical applications. Satisfactory performance has been achieved but often only in controlled environments. More recently, there has been increased demand for recognition of unconstrained face images, such as those collected from the internet [Huang et al., 2007] or captured by mobile devices and surveillance cameras [Beveridge et al., 2015]. However, recognition of unconstrained face images is a difficult problem due to degradation of face image quality and the wide variations of pose, illumination, expression, and occlusion often encountered in images [Ding and Tao, 2016].

The design of effective face image descriptors is a fundamental work for face recognition. A detailed summarization of existing face image descriptors can be found in Chapter 1. Despite the successful application of existing face image descriptors, the following three points are worth considering. First, the textual characteristics of human faces have mostly been overlooked in the design of existing descriptors. Second, it is generally prohibitive for hand-crafted descriptors to adopt a large sampling size due to the complication of the resulting encoding scheme and large feature size (i.e., the number of histogram bins) [Ul Hussain and Triggs, 2012]. However, a large sampling size is desirable since it provides better discriminative power, as has been proved by learning-based descriptors; it is therefore reasonable to ask whether it is genuinely impossible for hand-crafted descriptors to exploit large sampling sizes. Third, the recently proposed descriptors achieve good performance but at the cost of using computationally expensive techniques such as Gabor filtering and codebook learning. It would therefore be desirable to obtain a face image descriptor with superior performance that retains a lower computational cost and feature size.

To address these three limitations of existing techniques, in this chapter we present a novel face image descriptor named Dual-Cross Patterns (DCP). Inspired by the unique textural structure of human faces, DCP encodes second-order discriminative information in the directions of major facial components: the horizontal direction for eyes, eyebrows, and lips; the vertical direction for the nose; and the diagonal directions ($\pi/4$ and $3\pi/4$) for the end parts of facial components. The sampling strategy we adopt samples twice as many pixels as LBP. By appropriately grouping the sampled pixels

from the perspective of maximum joint Shannon entropy, we keep the DCP feature size reasonable. DCP is very efficient - only twice the computational cost of LBP. Significantly better performance is achieved even when a sub-DCP (denoted herein as DCP-1 and DCP-2) of exactly the same time and memory costs as LBP is used.

Furthermore, we propose a highly robust and discriminative face representation scheme called Multi-Directional Multi-Level Dual-Cross Patterns (MDML-DCPs). Specifically, the MDML-DCPs scheme employs the first derivative of Gaussian operator to conduct multi-directional filtering to reduce the impact of differences in illumination. DCP features are then computed at two levels: 1) holistic-level features incorporating facial contours and facial components and their configuration, and 2) component-level features focusing on the description of a single facial component. Thus, MDML-DCPs comprehensively encodes multi-level invariant characteristics of face images.

Both the DCP descriptor and the MDML-DCPs scheme are extensively evaluated on four large-scale databases: FERET [Phillips et al., 2000], CAS-PEAL-R1 [Gao et al., 2008], FRGC 2.0 [Phillips et al., 2005], and LFW [Huang et al., 2007]. The proposed DCP descriptor consistently achieves superior performance for both face identification and face verification tasks. More impressively, the proposed MDML-DCPs exploits only a single descriptor but achieves the best performance on two challenging unconstrained databases, FRGC 2.0 and LFW. Besides, this chapter provides a fair and systematic comparison between state-of-the-art facial descriptors, which has been rarely performed in the face recognition field [Akhtar et al., 2013].

The remainder of the chapter is organized as follows: Section 2.2 details the DCP descriptor and the construction of the MDML-DCPs face representation scheme is discussed in Section 2.3. Face recognition pipelines using MDML-DCPs are introduced in Section 2.4. Experimental results are presented in Section 2.5, leading to conclusions in Section 2.6.

2.2 Dual-Cross Patterns

The design of a face image descriptor consists of three main parts: image filtering, local sampling, and pattern encoding. The implementation of image filtering is flexible: possible methods include Gabor wavelets [Xie et al., 2010], Difference of Gaussian

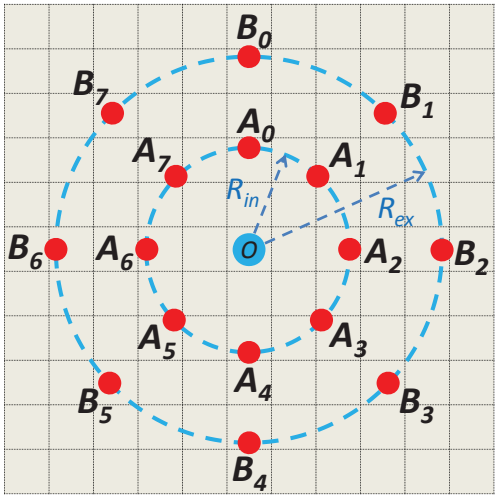


Figure 2.1: Local sampling of Dual-Cross Patterns. Sixteen points are sampled around the central pixel O. The sampled points A_0 to A_7 are uniformly spaced on an inner circle of radius R_{in}, while B_0 to B_7 are evenly distributed on the exterior circle with radius R_{ex}.

(DoG), or the recently proposed discriminative image filter [Lei et al., 2014]. In this chapter, we focus on local sampling and pattern encoding, which are the core components of a face image descriptor.

2.2.1 Local Sampling

The essence of DCP is to perform local sampling and pattern encoding in the most informative directions contained within face images. For face recognition, useful face image information consists of two parts: the configuration of facial components and the shape of each facial component. The shape of facial components is, in fact, rather regular. After geometric normalization of the face image, the central parts of several facial components, i.e., the eyebrows, eyes, nose, and mouth, extend either horizontally or vertically, while their ends converge in approximately diagonal directions ($\pi/4$ and $3\pi/4$). In addition, wrinkles in the forehead lie flat, while those in the cheeks are either raised or inclined.

Based on the above observations, local sampling of DCP is conducted as shown in Fig. 2.1. For each pixel O in the image, we symmetrically sample in the local neighborhood in the 0, $\pi/4$, $\pi/2$, $3\pi/4$, π, $5\pi/4$, $3\pi/2$, and $7\pi/4$ directions, which

are sufficient to summarize the extension directions of major facial textures. Two pixels are sampled in each direction. The resulting sampled points are denoted as $\{A_0, B_0; A_1, B_1; \cdots; A_7, B_7\}$. As illustrated in Fig. 2.1, $A_0, A_1, \cdots, A_7$ are uniformly spaced on an inner circle of radius R_{in}, while $B_0, B_1 \cdots, B_7$ are evenly distributed on the exterior circle with radius R_{ex}.

2.2.2 Pattern Encoding

Encoding of the sampled points is realized in two steps. First, textural information in each of the eight directions is independently encoded. Second, patterns in all eight directions are combined to form the DCP codes.

To quantize the textural information in each sampling direction, we assign each a unique decimal number:

$$DCP_i = S\left(I_{A_i} - I_O\right) \times 2 + S\left(I_{B_i} - I_{A_i}\right), 0 \leq i \leq 7, \tag{2.1}$$

where

$$S\left(x\right) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}, \tag{2.2}$$

and I_O, I_{A_i}, and I_{B_i} are the gray value of points O, A_i, and B_i, respectively. Therefore, four patterns are defined to encode the second-order statistics in each direction and each of the four patterns denotes one type of textural structure.

By simultaneously considering all eight directions, the total number of DCP codes is $4^8 = 65536$. This number is too large for practical face recognition applications; therefore, we adopt the following strategy. The eight directions are grouped into two subsets and each subset is further formulated as an encoder. In this way, the total number of local patterns is reduced to $4^4 \times 2 = 512$, which is computationally efficient. Although this strategy results in information loss, compactness and robustness of the descriptor are promoted. In the following subsection, we define the optimal grouping mode.

2.2.3 Dual-Cross Grouping

The grouping strategy introduced in the previous subsection produces $\frac{1}{2}\binom{8}{4} = 35$ combinations in total to partition all eight directions. To minimize information loss, we look for the optimal combination from the perspective of maximum joint Shannon entropy.

With the above analysis, $DCP_i\,(0 \leq i \leq 7)$ are discrete variables with four possible values: 0, 1, 2, and 3. Without loss of generality, the joint Shannon entropy for the subset $\{DCP_0, DCP_1, DCP_2, DCP_3\}$ is represented as

$$\begin{aligned} &H\left(DCP_0, DCP_1, DCP_2, DCP_3\right) \\ &= -\sum_{dcp_0} \cdots \sum_{dcp_3} P\left(dcp_0, \cdots, dcp_3\right) \log_2 P\left(dcp_0, \cdots, dcp_3\right), \end{aligned} \tag{2.3}$$

where dcp_0, dcp_1, dcp_2, and dcp_3 are particular values of DCP_0, DCP_1, DCP_2, and DCP_3, respectively. And $P\left(dcp_0, \cdots, dcp_3\right)$ is the probability of these values occurring simultaneously. The maximum joint Shannon entropy of the four variables is achieved when they are statistically independent.

In real images, the more sparsely the pixels are scattered, the more independent they are. Therefore, the maximum joint Shannon entropy for each subset is achieved when the distance between the sampled points is at its maximum. As a result, we define $\{DCP_0, DCP_2, DCP_4, DCP_6\}$ as the first subset and $\{DCP_1, DCP_3, DCP_5, DCP_7\}$ as the second subset. The resulting two subsets are illustrated in Fig. 2.2. Since each of the two subsets constructs the shape of a cross, the proposed descriptor is named Dual-Cross Patterns. In Section 2.5, we empirically validate that dual-cross grouping achieves the maximum joint Shannon entropy among all grouping modes on the FERET database.

2.2.4 DCP Face Image Descriptor

We name the two cross encoders DCP-1 and DCP-2, respectively. The codes produced by the two encoders at each pixel O are represented as

$$DCP\text{-}1 = \sum_{i=0}^{3} DCP_{2i} \times 4^i, \tag{2.4}$$

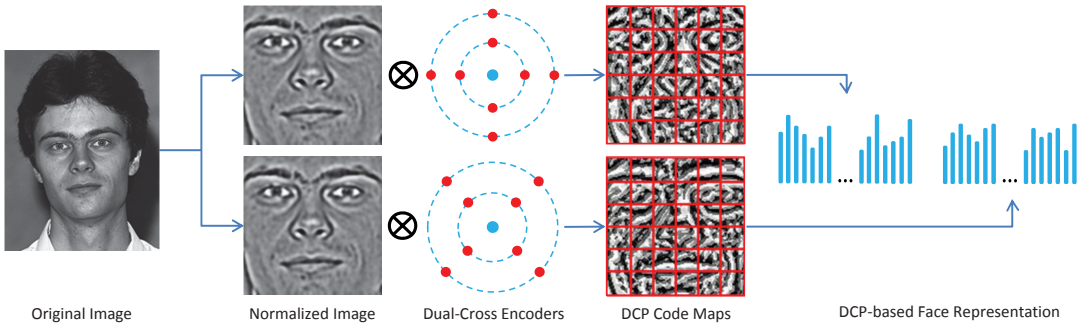


Figure 2.2: Face representation using Dual-Cross Patterns. The normalized face image is encoded by the two cross encoders, respectively. Concatenation of the regional DCP code histograms forms the DCP-based face representation.

$$DCP\text{-}2 = \sum_{i=0}^{3} DCP_{2i+1} \times 4^i. \tag{2.5}$$

The DCP descriptor for each pixel O in an image is the concatenation of the two codes generated by the two cross encoders:

$$DCP = \left\{ \sum_{i=0}^{3} DCP_{2i} \times 4^i, \sum_{i=0}^{3} DCP_{2i+1} \times 4^i \right\}. \tag{2.6}$$

After encoding each pixel in the face image using the dual-cross encoders, two code maps are produced that are respectively divided into a grid of non-overlapping regions. Histograms of DCP codes are computed in each region and all histograms are concatenated to form the holistic face representation. The overall framework of the above face representation approach is illustrated in Fig. 2.2. This face representation can be directly used to measure the similarity between a pair of face images using metrics such as the chi-squared distance or histogram intersection. The computation of the DCP descriptor, which doubles the feature size of LBP, is very efficient by only doubling the time cost of LBP.

We notice that two recently proposed descriptors DFD [Lei et al., 2014] and Center Symmetric-Pairs of Pixels (CCS-POP) [Choi et al., 2012] adopt similar sampling modes to DCP. However, they are essentially different from DCP. First, there is no clear motivation for the two descriptors that why such a sampling mode is suitable for face

images. Second, the pattern encoding strategies are different: both the two descriptors rely on learning algorithms to handle the large sampling size problem mentioned in Section 2.1.

2.3 Multi-Directional Multi-Level Dual-Cross Patterns

We now present a face representation scheme based on DCP named Multi-Directional Multi-Level Dual-Cross Patterns (MDML-DCPs) to explicitly handle the challenges encountered in unconstrained face recognition.

2.3.1 The MDML-DCPs Scheme

A major difficulty in unconstrained face recognition is that many factors produce significant intra-personal differences in the appearance of face images, in particular variations in illumination, image blur, occlusion, and pose and expression changes. We mitigate the influence of these factors using multi-directional gradient filtering and multi-level face representation.

In MDML-DCPs, the first derivative of Gaussian operator (FDG) is exploited to convert a gray-scale face image into multi-directional gradient images that are more robust to variations in illumination. The FDG gradient filter of orientation θ can be expressed as follows:

$$FDG\left(\theta\right) = \frac{\partial G}{\partial \mathbf{n}} = \mathbf{n} \cdot \nabla G, \tag{2.7}$$

where $\mathbf{n} = (cos\theta, sin\theta)$ is the normal vector standing for the filtering direction and $G = \exp\left(-\frac{x^2+y^2}{\sigma^2}\right)$ is a two-dimensional Gaussian filter. The application of FDG is inspired by the classical work of Canny [Canny, 1986], which proved that FDG is the optimal gradient filter according to three criteria, namely signal-to-noise ratio (SNR) maximization, edge location accuracy preservation, and single response to single edge. These three criteria are also relevant for face recognition, where it is desirable to enhance the facial textural information while suppressing noise. FDG significantly saves computational cost compared to other gradient-like filters, such as Gabor wavelets [Xie et al., 2010]. We denote the concatenation of DCP descriptors extracted from the FDG-filtered images as MD-DCPs.

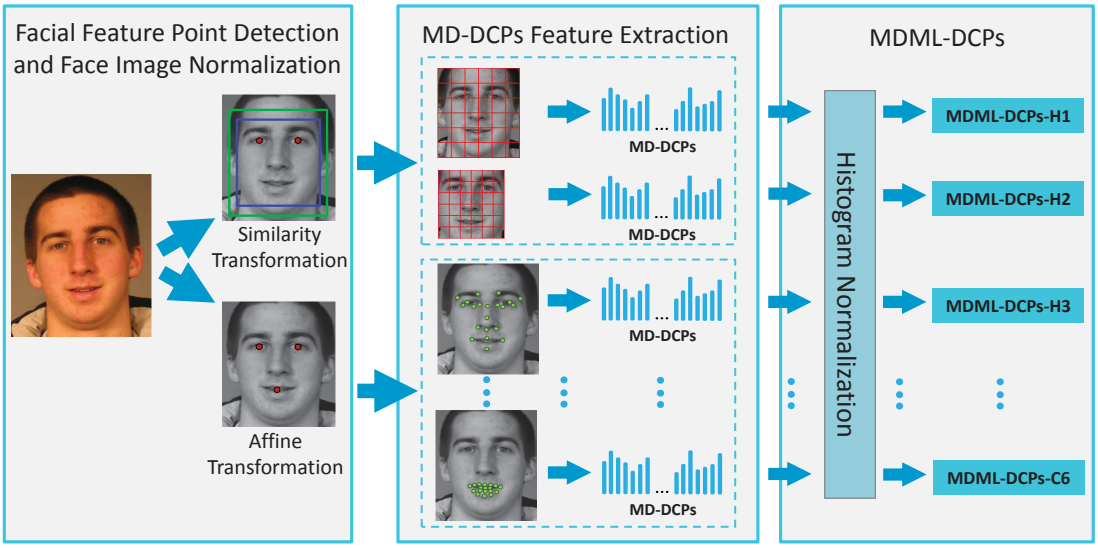


Figure 2.3: Framework of the MDML-DCPs face representation scheme. MDML-DCPs-H1 and MDML-DCPs-H2 are extracted from the rectified image by similarity transformation. MDML-DCPs-H3, MDML-DCPs-C1 to C6 are extracted from the affine-transformed image. The MDML-DCPs face representation is the set of the above nine feature vectors.

To build pose-robust face representation, MDML-DCPs normalizes the face image by two geometric rectifications based on a similarity transformation and an affine transformation, respectively. The similarity transformation retains the original information of facial contours and facial components and their configuration. The affine transformation reduces differences in intra-personal appearance caused by pose variation.

MDML-DCPs combines both holistic-level and component-level features, which are computed on the normalized images by the two transformations. Holistic-level features capture comprehensive information on both facial components and facial contour. However, it is also sensitive to changes in appearance of each component caused by occlusion, pose, and variations in expression. In contrast, component-level features focus on the description of a single facial component, and thus are independent of changes in appearance of the other components. In this way the information generated by these two feature levels is complementary and appropriate fusion of the two promotes robustness to interference.

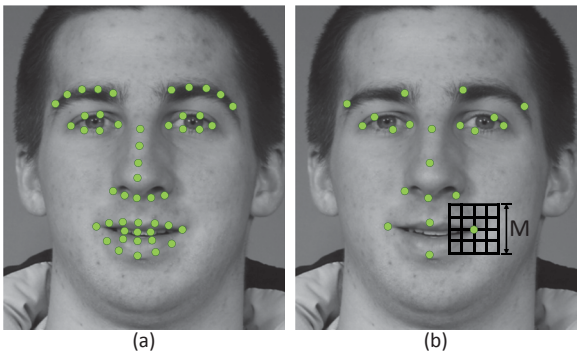


Figure 2.4: (a) The 49 facial feature points detected by the face alignment algorithm. (b) MDML-DCPs-H3 employs 21 facial feature points over all facial components. MDML-DCPs-C1 to C6 respectively select 10 facial feature points on both eyebrows, 12 points on both eyes, 11 points on the left eye and left eyebrow, 11 points on the right eye and right eyebrow, 9 points on nose, and 18 points on mouth. Around each facial feature point, MD-DCPs are extracted from $J \times J$ (in this figure, $J = 4$) non-overlapping regions within the patch of size $M \times M$ pixels.

2.3.2 Implementation Details

Similar to most state-of-the-art face representation schemes, e.g., [Chen et al., 2013, Taigman et al., 2014a], MDML-DCPs also benefits from recent progress in face alignment algorithms that accurately locate dense facial feature points in real-time. The MDML-DCPs face representation scheme is shown in Fig. 2.3. In MDML-DCPs, a high performance face alignment algorithm based on [Xiong and De la Torre, 2013] is first employed to locate 49 facial feature points (as shown in Fig. 2.4a), before applying the two geometric rectifications.

The similarity transformation is based on the two detected eye centers. In the rectified image, we compute MD-DCPs at two holistic levels: 1) MD-DCPs of non-overlapping regions of the external cropped face image (including facial contour), and 2) MD-DCPs of non-overlapping regions of the internal cropped face image (without facial contour). The first feature encodes both facial contour and facial components while the second feature focuses on encoding facial components only, which are free from background interference. For clarity, we denote the two features as MDML-DCPs-H1 and MDML-DCPs-H2.

The affine transformation is determined by the three detected facial feature points: the centers of the two eyes and the center of the mouth. In the rectified image,

one holistic-level feature denoted as MDML-DCPs-H3, and six component-level features referred to MDML-DCPs-C1 to MDML-DCPs-C6, are computed based on the detected dense facial feature points. As shown in Fig. 2.4b, the method for feature extraction around each facial feature point is similar to the approaches used in [Chen et al., 2013, Prince et al., 2008]: centered on each facial feature point, a patch of size $M \times M$ pixels is located and further divided into $J \times J$ non-overlapping regions. The concatenated MD-DCPs feature of the J^2 regions forms the description of the feature point. MDML-DCPs-H3 and MDML-DCPs-C1 to C6 are formed by respectively concatenating the descriptions of different facial feature points. As shown in Fig. 2.4, MDML-DCPs-H3 selects 21 facial feature points over all facial components, while MDML-DCPs-C1 to C6 select feature points on only one particular facial component. Elements of the three holistic-level features and six component-level features are normalized by the square root. Together, the set of the nine normalized feature vectors form the MDML-DCPs face representation.

2.4 Face Recognition Algorithm

In this section, the face matching problem is addressed using the proposed MDML-DCPs face representation scheme. First, one classifier is built for each of the nine feature vectors. Then, the similarity scores of the nine classifiers are fused by linear SVM or simple averaging.

Two algorithms are considered: Whitened Principal Component Analysis (WPCA; an unsupervised learning algorithm) and Probabilistic Linear Discriminant Analysis (PLDA; a supervised learning algorithm) [Li et al., 2012b, Prince and Elder, 2007]. The choice of which of the two algorithms to use is dataset dependent: for datasets that have a training set with multiple face images for each subject, we choose PLDA; otherwise, WPCA is used.

2.4.1 WPCA

Principal Component Analysis (PCA) learns an orthogonal projection matrix U from training data and projects high-dimensional feature vector x to the low-dimensional

vector y,

$$y = U^T x. \tag{2.8}$$

The columns of U are composed of the leading eigenvectors of the covariance matrix of the training data. However, the first few eigenvectors in U encode mostly variations in illumination and expression, rather than information that discriminates identity. The whitening transformation tackles this problem by normalizing the contribution of each principal component

$$y = \left(U\Lambda^{-1/2}\right)^T x, \tag{2.9}$$

where $\Lambda = diag\left\{\lambda_1, \lambda_2, \cdots\right\}$ with λ_i being the ith leading eigenvalue. After projecting the facial feature vectors to the low-dimensional subspace using WPCA, the similarity score between two feature vectors y_1 and y_2 is measured by the cosine metric

$$Sim\left(y_1, y_2\right) = \frac{y_1^T y_2}{\|y_1\| \, \|y_2\|}. \tag{2.10}$$

2.4.2 PCA combined with PLDA

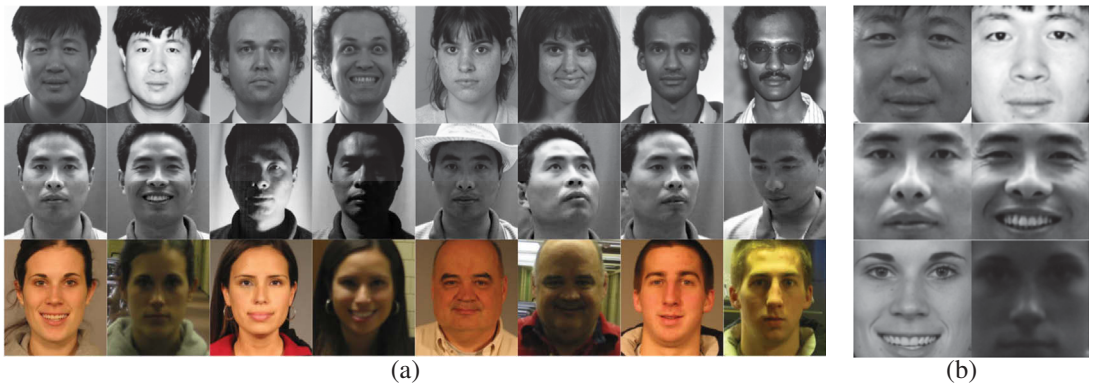
The feature vectors in MDML-DCPs are high-dimensional. To effectively apply PLDA, the dimensionality of the nine feature vectors is first reduced by PCA.

PLDA models the face data generation process

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \varepsilon_{ij}, \tag{2.11}$$

where x_{ij} denotes the jth face data of the ith individual. μ is the mean of all face data. F and G are factor matrices whose columns are the basis vectors of the between-individual subspace and the within-individual subspace, respectively. h_i is the latent identity variable that is constant for all images of the ith subject. w_{ij} and ε_{ij} are noise terms explaining intra-personal variance [Li et al., 2012b].

It is shown in [Li et al., 2012b] that the identification and verification problems can be solved by computing the log-likelihood ratio that whether two observed images share the same identity variable h or not. In this chapter, we refer to the log-likelihood ratio as similarity score for consistency with the case of the WPCA classifier.



(a) (b)

Figure 2.5: (a) Sample images from FERET (first row), CAS-PEAL-R1 (second row) and FRGC 2.0 (third row) containing typical variations in each database. (b) Samples of normalized images of size 128×128 pixels.

Figure 2.6: Sample images from LFW. Images in the two rows are aligned by a similarity transformation and an affine transformation, respectively.

2.5 Experiments

In this section, the proposed DCP and MDML-DCPs are extensively evaluated in both face identification and face verification tasks. Experiments are conducted on four publicly available large-scale face databases: FERET, CAS-PEAL-R1, FRGC 2.0, and LFW. Example images of the four databases are shown in Figs. 2.5 and 2.6.

The FERET database contains one gallery set Fa and four probe sets, i.e., Fb, Fc, Dup1, and Dup2. In this chapter, the standard face identification protocol specified in [Phillips et al., 2000] is employed.

The CAS-PEAL-R1 database includes one training set, one gallery set, and nine probe sets. Each of the nine probe sets is restricted to one type of variations. In detail,

the PE, PA, PL, PT, PB, and PS probe sets correspond to variations in expression, accessory, lighting, time, background, and distance of frontal faces, respectively. The PU, PM, and PD probe sets correspond to one type of pose variation. The standard identification protocol defined in [Gao et al., 2008] is followed.

The FRGC 2.0 database incorporates data for six face verification experiments [Phillips et al., 2005]. We focus on Experiments 1 and 4. The shared target set of Experiments 1 and 4 is collected from frontal facial images taken under controlled illumination, while images in their query sets are captured under controlled and uncontrolled conditions, respectively. Verification rates at 0.1% FAR are reported.

The LFW database [Huang et al., 2007] contains 13,233 unconstrained face images that are organized into two "Views". View 1 is for model selection and parameter tuning while View 2 is for performance reporting. Two paradigms are used to exploit the training set in View 2, an image-restricted paradigm and an image-unrestricted paradigm. In the first paradigm, only the officially defined image pairs are available for training. In the second paradigm, the identity information of the training images can be used. We report the mean verification accuracy and standard error of the mean (S_E) on the View 2 data.

Five experiments are conducted. First, the dual-cross grouping mode for DCP is empirically proved to achieve the maximum joint Shannon entropy. Second, parameter selection of DCP is described. Third, the performance of DCP is compared with eleven state-of-the-art face image descriptors. Fourth, the performance of MD-DCPs is evaluated to determine the contribution of multi-directional filtering by FDG. Finally, the power of the MDML-DCPs face representation scheme is presented. More experimental results, e.g., parameter selection of DCP, are available in the appendix.

In the first four experiments, we focus on the evaluation of face image descriptors. In these four experiments, all face images are cropped and resized to 128×128 ($rows \times columns$) pixels with the eye centers located at $(34, 31)$ and $(34, 98)$, as shown in Fig. 2.5b. For LFW, we use the aligned version (LFW-A) [Wolf et al., 2011b], while for the other three databases, images are cropped according to the eye coordinates provided by the databases. The cropped images are photometrically normalized using a simple operator (denoted as TT) developed by Tan & Triggs [Tan and Triggs, 2010] and then encoded by one face image descriptor. Each encoded face image is divided into $N \times N$ non-overlapping regions. Concatenating the code histograms of these

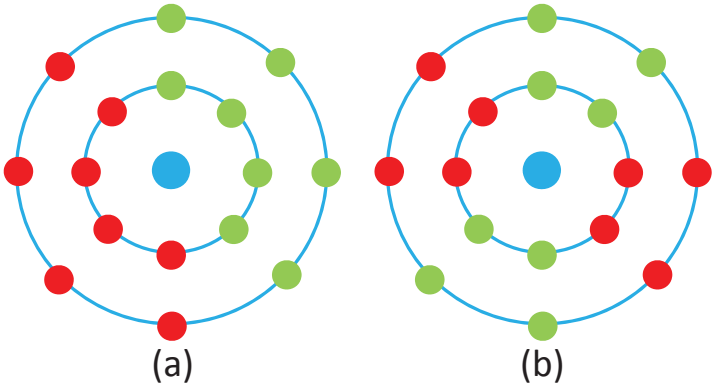


Figure 2.7: Another two representative grouping modes for the eight sampling directions of DCP. Sampled points of the same colour belong to the same subset.

regions forms the representation for the image.

In the fifth experiment, we aim to demonstrate that the proposed MDML-DCPs face representation scheme has excellent performance. In this experiment, all face images are normalized by two geometric transformations, as illustrated in Fig. 2.6. In both transformations, face images are resampled to 180×162 pixels with the eye centers mapped to $(66, 59)$ and $(66, 103)$. For the affine-transformed face images, the centers of the mouths are unified to $(116, 81)$. TT is applied to the resampled images for photometric normalization.

2.5.1 Empirical Justification for Dual-Cross Grouping

Section 2.2 intuitively suggests that the dual-cross grouping mode is optimal from the perspective of the joint Shannon entropy maximization. In this experiment, we empirically validate this point on the FERET database. For each of the 35 possible grouping modes, $DCP_i\,(0 \leq i \leq 7)$ are divided into two subsets. Then, the joint Shannon entropy for each of the two subsets are calculated on one image using (2.3) and are summed together. The above process is repeated on the 1,196 gallery images of the FERET database. The mean value of the summed joint Shannon entropy is recorded.

Experimental results show that dual-cross grouping mode achieves the highest joint Shannon entropy among all 35 grouping modes. Fig. 2.8 characterizes the joint Shannon entropy as a function of R_{in} and R_{ex} and illustrate the superiority of

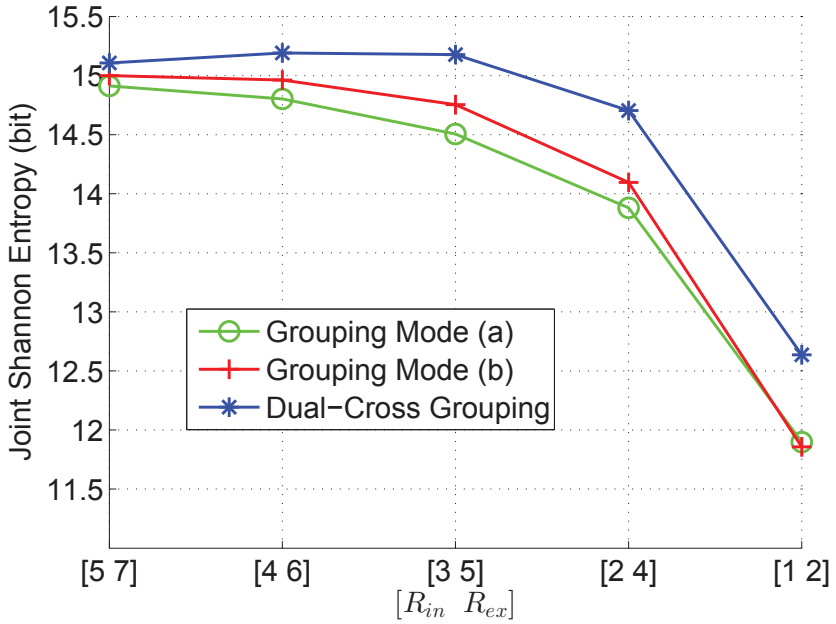


Figure 2.8: Joint Shannon entropy as a function of R_{in} and R_{ex}. Three grouping modes are evaluated in this figure: modes (a) and (b) in Fig. 2.7 and the dual-cross grouping.

dual-cross grouping mode by comparing it with two representative grouping modes, exampled in Fig. 2.7. Note that the joint Shannon entropy is related to the sampling radii of DCP: smaller sampling radii mean stronger dependence among the sampled points, which results in a smaller joint Shannon entropy. The dual-cross grouping mode achieves the highest entropy under all sets of the radii values. Therefore, the dual-cross grouping mode is empirically optimal.

2.5.2 Parameter Selection of DCP

In this experiment, we take the FERET database [Phillips et al., 2000] for example to illustrate the influence of the DCP parameters on its performance. The face images are normalized to 128×128 pixels, as described in Section 2.5 of the chapter. There are three parameters in DCP, the radii R_{in} and R_{ex} (with different values of R_{in} and R_{ex} capturing information on different scales) and the region number N. A larger value of N helps to preserve spatial information but makes the descriptor more sensitive to misalignment errors. The chi-squared metric is used to measure the similarity of two face images. In this experiment, LBP is also evaluated for comparison with a

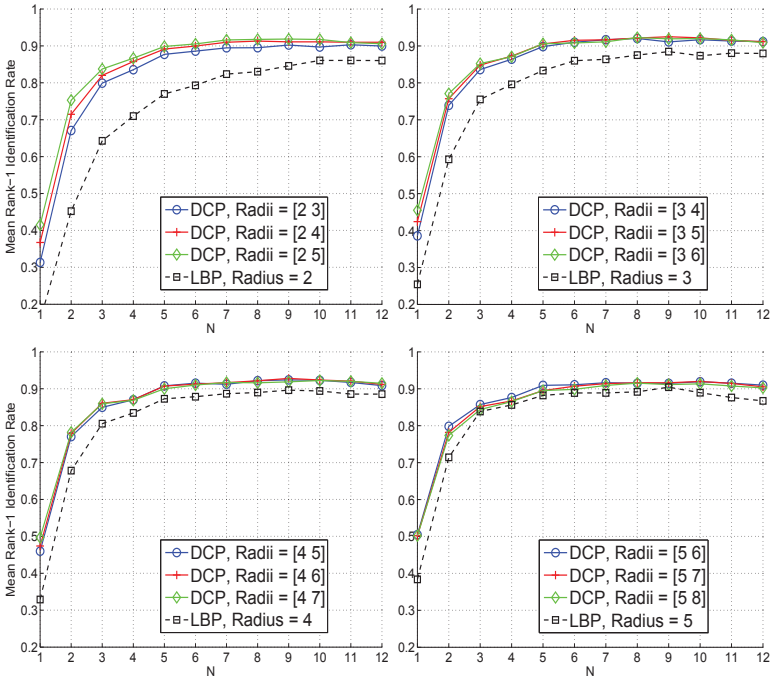


Figure 2.9: The mean rank-1 identification rates of DCP and LBP on four FERET probe sets as a function of N.

sampled points number of eight and exploiting all 256 LBP codes. The radius of LBP is restricted to the value of R_{in} to highlight the importance of including more sampling points for effective pattern encoding.

The mean rank-1 identification rates against N for the four FERET probe sets using DCP and LBP are plotted in Fig. 2.9. Optimal performance of DCP is achieved when $N = 9$, $R_{in} = 4$ and $R_{ex} = 6$. DCP consistently achieves better performance than LBP with all sets of parameters. In particular, DCP outperforms LBP by 10% to 30% when $N \leq 2$, which justifies the motivation of DCP to include more sampling points for pattern encoding.

The chi-squared metric has been used to measure the similarity of two face images, but it is noted that DCP works equally well when assessed using the histogram intersection metric. The mean identification rates using DCP measured with the chi-squared and histogram intersection metrics are 92.80% and 92.66%, respectively.

2.5.3 Evaluation of the Performance of DCP

In this subsection, the performance of DCP, DCP-1, and DCP-2 for both face identification and face verification tasks are evaluated. To better illustrate the advantages of DCP, the performance of eleven state-of-the-art face image descriptors, i.e., LBP [Ahonen et al., 2006], LTP [Tan and Triggs, 2010], LPQ [Ahonen et al., 2008], POEM [Vu and Caplier, 2012], Local Gabor XOR Patterns (LGXP) [Xie et al., 2010], Multi-scale LBP (MsLBP), Multi-scale tLBP (MsTLBP) [Trefnỳ and Matas, 2010], Multi-scale dLBP (MsDLBP) [Trefnỳ and Matas, 2010], LQP [Hussain et al., 2012], DFD [Lei et al., 2014], and CHG [Choi et al., 2012] are also presented. [1]

The first eight descriptors are hand-crafted and they are easy to implement. We therefore test them together with DCP. It is worth noting that the cropped face image data for these descriptors are exactly the same. All of them extract features from photometrically normalized images by TT. The parameters for each of the descriptors and TT are carefully tuned on each database and different distance metrics (chi-squared or histogram intersection) are tested. Finally, the best results for each descriptor are reported. Therefore, the experimental results directly compare the recognition capabilities of the descriptors.

On the other hand, LQP, DFD, and CHG are learning-based descriptors, which are complicated to implement. In this chapter, their performance on FERET and LFW is directly cited from the original papers, even though the experimental settings are different from that of DCP. For LQP, we cite its performance with the image filtering step by Gabor wavelets, which is more robust to illumination variation. Before presenting the detailed experimental results, the feature size (number of histogram bins) for each descriptor except CHG (CHG is not a histogram-based descriptor) is listed in Table 2.1. In the following experiments, LBP-based descriptors are implemented without using uniform coding [Ojala et al., 2002] (using uniform coding actually degrades the performance of LBP, LTP, and POEM, et al.). For POEM, we compute LBP codes on four-orientation gradient magnitude maps, so its feature size is 1024 in this chapter.

[1] For Table 2.2 to Table 2.6, a superscript * means that the results are cited from the original papers. A suffix '-Flip' means that the descriptor adopts the 'flip' trick [Hussain et al., 2012].

Table 2.1: Feature Size of the Investigated Face Image Descriptors

Descriptor	Feature Size	Descriptor	Feature Size
LBP [Ahonen et al., 2006]	256	MsDLBP [Trefnỳ and Matas, 2010]	512
MsLBP [Ojala et al., 2002]	512	LQP [Hussain et al., 2012]	300
LTP [Tan and Triggs, 2010]	512	DFD [Lei et al., 2014]	1024
LPQ [Ahonen et al., 2008]	256	DCP-1	256
POEM [Vu and Caplier, 2012]	1024	DCP-2	256
LGXP [Xie et al., 2010]	640	DCP	512
MsTLBP [Trefnỳ and Matas, 2010]	512		

2.5.3.1 Face Identification: FERET

Face identification experiments are conducted on the FERET and CAS-PEAL-R1 databases. The rank-1 identification rates on the four probe sets of FERET are presented in Table 2.2. We make four observations:

- While high performance is achieved by all descriptors on the well controlled Fb and Fc probe sets, DCP still outperforms MsLBP by over 1%. In particular, DCP achieves a perfect identification rate on the Fc set.
- There is a substantial performance drop for all descriptors on Dup1 and Dup2 probe sets, in which images contain moderate expression and illumination variations. DCP performs best on both sets, with a margin of 1.28% on Dup2 by comparing with the second best descriptor. The performance of MsTLBP ranks second on Dup1 and LGXP ranks second on Dup2. It is worth pointing out that LGXP depends on 80 expensive convolutions and produces a larger feature size, while both DCP and MsTLBP have low computational complexity.
- Both DCP-1 and DCP-2 perform better than most of the other descriptors at lower computational cost. Their time and memory costs are exactly the same as those of LBP, suggesting the sampling and encoding strategies of DCP are highly effective.
- As expected, the performance of DCP is better than both DCP-1 and DCP-2, which means that DCP-1 and DCP-2 contain complementary information. When DCP-1 and DCP-2 are combined, the mean identification rate increases by over 1%.

2.5.3.2 Face Identification: CAS-PEAL-R1

Identification results on the nine probe sets of CAS-PEAL-R1 are shown in Table 2.3. For each descriptor, results are reported with the set of parameters that achieves the highest mean identification rate over all nine probe sets. We make the following observations:

Table 2.2: Identification Rates for Different Descriptors on FERET

	Fb	Fc	Dup1	Dup2	Mean
LBP	96.90	98.45	83.93	82.48	90.44
LTP	96.90	98.97	83.93	83.76	90.89
LPQ	97.41	99.48	82.69	81.62	90.30
POEM	98.24	99.48	82.83	82.05	90.65
LGXP	97.32	99.48	85.46	85.47	91.93
MsLBP	97.07	98.97	83.38	83.33	90.69
MsTLBP	98.16	99.48	85.73	85.04	92.10
MsDLBP	95.65	97.94	79.09	79.49	88.04
LQP-Flip*	**99.50**	99.50	81.20	79.90	90.03
DFD*	99.20	98.50	85.00	82.90	91.40
CHG*	97.50	98.50	85.60	84.60	91.55
DCP-1	97.91	98.45	84.49	84.19	91.26
DCP-2	97.99	99.48	84.35	85.04	91.72
DCP	98.16	**100.0**	**86.29**	**86.75**	**92.80**

- In general, the performance of all the descriptors is good on the PE, PA, PT, PB, and PS probe sets, with DCP producing the highest mean identification rate and MsTLBP the second highest. The performance of all descriptors is poor on the remaining PL, PU, PM, and PD probe sets due to serious illumination and pose variations.

- DCP outperforms MsLBP and LTP by 2.5% and 3.08% on the PL set, respectively. However, the performance of DCP is lower than LPQ, POEM, and LGXP. The PL images contain not only rich illumination variation, but also serious image blur, which explains the superior performance of the blur-invariant descriptor LPQ. LGXP and POEM benefit from image filtering steps that turn the gray-scale image into gradient images. We show that by introducing multi-directional gradient filtering by FDG, the performance of DCP is significantly improved.

- DCP exhibits excellent robustness to pose variations on the PU, PM, and PD probe sets, with superior mean recognition rate by as much as 3.44%. This result is encouraging, since pose variation is a major challenge in unconstrained face recognition [Ding and Tao, 2016].

2.5.3.3 Face Verification: FRGC 2.0

Face verification experiments are conducted on the FRGC 2.0 and LFW databases. Verification rates at 0.1% FAR on Experiments 1 and 4 of FRGC 2.0 are listed in Tables 2.4 and 2.5, respectively. As mentioned above, the query sets of the two experiments are composed of controlled and uncontrolled images, respectively. The query images in Experiment 4 are degraded by serious image blur and significant illumination variation, making this dataset very challenging. We make the following observations:

- In Experiment 1, DCP shows the best performance. Both DCP-1 and DCP-2 achieve better performance than most of the existing descriptors. These observations are consistent with the results on FERET.

- In Experiment 4, the mean verification rate of DCP is higher than MsLBP and LTP by 2.64% and 1.09%, respectively. Due to the lack of an image filtering step in DCP, its performance is lower than that of LGXP and POEM. This result is consistent with that seen for the PL set of CAS-PEAL-R1, whose images also feature serious illumination variation and image blur.

2.5.3.4 Face Verification: LFW

LFW is a very challenging dataset since its images contain large pose, expression, and illumination variations. Experiments in this section follow the image-restricted paradigm. For each descriptor, its parameters are tuned on the View 1 data, and performance on the View 2 data is reported in Table 2.6. We observe that:

- Among the manually-designed descriptors, DCP achieves the best performance while DCP-1 ranks second. This result is consistent with the results on FERET

Table 2.3: Rank-1 Identification Rates for Different Face Image Descriptors on the Nine Probe Sets of PEAL

	PE	PA	PT	PB	PS	PL	PU	PM	PD	Mean (PE-PS)	Mean (PE-PL)	Mean (PU-PD)
LBP	94.27	91.82	100.0	**99.46**	**99.64**	46.90	60.32	83.66	44.60	97.04	88.68	62.86
LTP	94.39	91.77	100.0	**99.46**	**99.64**	47.17	61.32	84.46	44.68	97.05	88.74	63.49
LPQ	93.95	92.39	100.0	99.28	99.27	57.16	57.78	86.46	44.76	96.98	90.34	63.00
POEM	95.54	92.39	100.0	**99.46**	**99.64**	54.66	58.14	85.12	42.52	97.41	90.28	61.93
LGXP	94.97	91.33	100.0	99.28	**99.64**	**63.26**	39.46	66.83	22.91	97.04	**91.41**	43.07
MsLBP	95.16	92.04	100.0	**99.46**	**99.64**	47.75	61.76	85.16	44.88	97.26	89.01	63.93
MsTLBP	95.41	92.74	100.0	**99.46**	**99.64**	48.06	60.98	87.34	45.48	97.45	89.22	64.60
MsDLBP	92.42	90.63	100.0	99.28	99.27	48.11	55.00	82.03	37.03	96.32	88.29	58.02
DFD*	**99.30**	**94.40**	-	-	-	59.00	-	-	-	-	-	-
DCP-1	95.99	92.60	100.0	98.73	99.27	46.37	60.68	85.74	48.14	97.32	88.83	64.85
DCP-2	95.54	91.90	100.0	98.92	**99.64**	43.91	60.64	83.08	43.62	97.20	88.32	62.45
DCP	96.11	92.82	100.0	99.10	**99.64**	50.25	**65.39**	**87.44**	**51.30**	**97.53**	89.65	**68.04**

Table 2.4: Verification Results on the FRGC 2.0 Experiment 1

	ROC I	ROC II	ROC III	Mean
LBP	89.93	87.53	84.83	87.43
LTP	90.98	88.63	86.07	88.56
LPQ	89.96	87.38	84.48	87.27
POEM	90.60	88.26	85.63	88.16
LGXP	91.89	88.80	85.33	88.67
MsLBP	90.78	88.45	85.87	88.37
MsTLBP	**93.24**	91.16	88.83	91.08
MsDLBP	87.65	84.93	81.96	84.85
DCP-1	91.40	89.14	86.65	89.06
DCP-2	92.96	90.78	88.42	90.72
DCP	93.22	**91.21**	**88.93**	**91.12**

Table 2.5: Verification Results on the FRGC 2.0 Experiment 4

	ROC I	ROC II	ROC III	Mean
LBP	18.62	19.12	20.07	19.27
LTP	20.21	21.39	22.99	21.53
LPQ	22.53	23.25	24.39	23.39
POEM	29.99	30.60	**31.46**	30.68
LGXP	**32.44**	**31.81**	31.13	**31.79**
MsLBP	19.34	19.80	20.81	19.98
MsTLBP	20.04	20.25	21.02	20.44
MsDLBP	18.34	18.85	19.60	18.93
DCP-1	20.10	20.57	21.51	20.73
DCP-2	18.22	18.74	19.58	18.85
DCP	21.78	22.49	23.59	22.62

and Experiment 1 on FRGC 2.0. The robustness of DCP to pose variations is consistent with the results on CAS-PEAL-R1.

Table 2.6: Mean Verification Accuracy on the LFW View 2 Data

	Accuracy(%)$\pm S_E$		Accuracy(%)$\pm S_E$
LBP	72.43 $\pm$ 0.49	MsDLBP	72.17 $\pm$ 0.59
LTP	72.65 $\pm$ 0.52	LQP-Flip*	75.30 $\pm$ 0.26
LPQ	72.68 $\pm$ 0.46	**DFD-Flip***	**80.02 $\pm$ 0.50**
POEM	73.98 $\pm$ 0.56	DCP-1	74.50 $\pm$ 0.48
LGXP	70.58 $\pm$ 0.43	DCP-2	73.28 $\pm$ 0.48
MsLBP	72.88 $\pm$ 0.50	DCP	75.00 $\pm$ 0.64
MsTLBP	74.12 $\pm$ 0.57	DCP-Flip	76.37 $\pm$ 0.71

- LGXP does not perform well on LFW because LFW images contain significant pose variations. Although LGXP is robust to serious illumination variations, it is sensitive to pose variations, consistent with the results seen from experimentation on the non-frontal probe sets of CAS-PEAL-R1.

- Compared with the two learning-based descriptors, the performance of DCP is better than that of LQP, but is worse than that of DFD. However, DCP has clear advantage in both time and memory costs. Besides, DFD adopts supervised learning algorithms to enhance its discriminative power, while DCP is independent of any learning algorithms. We show that by applying supervised learning algorithms to the extracted DCP feature, superior performance can be achieved.

2.5.3.5 Discussion

The above experimental results reveal that DCP performs quite well across a range of evaluations. In particular, DCP significantly outperforms MsLBP, which is of exactly the same time and memory costs as DCP, as shown in Fig. 2.10. The performance of MsTLBP is usually slightly worse than DCP on well-controlled datasets. However, DCP considerably outperforms MsTLBP on the challenging PU and PD probe sets of CAS-PEAL-R1, and the Dup2 probe set of FERET. This indicates DCP is a more robust descriptor to pose and aging factors.

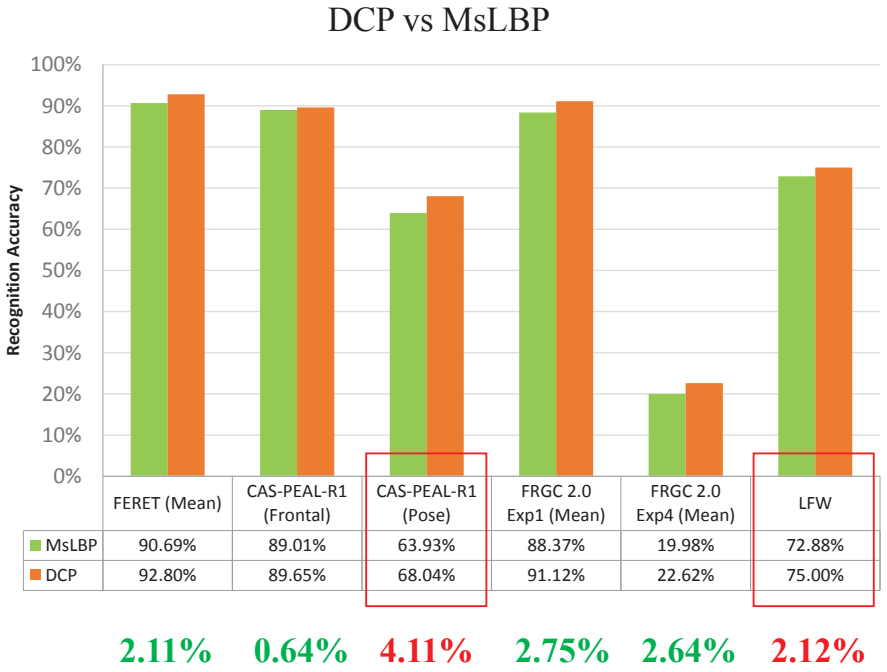


Figure 2.10: Performance comparison between DCP and MsLBP on the four face datasets.

The excellent performance of DCP can be explained as follows. It has large sampling size to encode more discriminative information; it encodes the second-order statistics in the most informative directions on human face; and the dual-cross grouping strategy ensures good complementarity between the divided two encoders.

2.5.4 The Contribution of Multi-directional Filtering

DCP has been shown to have strong discriminative power and excellent robustness to expression, pose, and moderate illumination variations. It has also been shown that when there is serious illumination variation and image blur, the performance of DCP is degraded due to the lack of an image filtering step. In this section, we show that MD-DCPs that extracts the DCP feature on FDG-filtered images is robust to illumination variation. The roles of FDG are two-fold: first, it converts the face image into gradient maps, which are more robust to illumination variation, and second, FDG is proven to achieve high SNR [Canny, 1986], which is vital for the performance of face image descriptors for low quality images.

Throughout this chapter, the number of filtering orientations by FDG is fixed to 4 ($\theta = 0, \pi/4, \pi/2,$ and $3\pi/4$) and the parameter σ of the Gaussian kernel is set at 1.0. Experiments are conducted on three datasets:

- A new FERET probe set is constructed by collecting the 146 most challenging FERET probe images. The FERET Fa set is still utilized as the gallery set. The identification rate of DCP on this probe set is only 17.12%. In contrast, the identification rate when MD-DCPs is used increases to 33.56%.
- For the PL probe set of CAS-PEAL-R1, the identification rates of DCP and MD-DCPs are 50.25% and 65.23%, respectively. When considered along with the previously best results shown in Table 2.3, it is evident that the performance of MD-DCPs is superior to that of LPQ (57.16%), POEM (54.66%), and LGXP (63.26%).
- For the face verification task on FRGC 2.0 Experiment 4, the mean verification rate of MD-DCPs at 0.1% FAR for ROC I, ROC II, and ROC III is 30.74%. Together with the previously best results shown in Table 2.5, it is clear that MD-DCPs outperforms LPQ (23.39%) and POEM (30.68%). LGXP outperforms MD-DCPs with the mean verification rate of 31.79%, but at the cost of 80 expensive convolution operations.

In conclusion, MD-DCPs consistently achieves excellent performance, which means that multi-directional filtering by FDG is indeed helpful for removing illumination variation and enhancing the robustness of face image descriptors on low SNR images.

2.5.5 Performance Evaluation of MDML-DCPs

Before presenting the performance of MDML-DCPs, we first unify most of its parameters for testing the four databases. For the holistic-level feature MDML-DCPs-H1, we compute it within the area whose top left corner is located at $(33, 27)$, while the bottom right corner is located at $(154, 136)$ in the normalized image by similarity transformation. The holistic-level feature MDML-DCPs-H2 is computed within the area where the two corners are located at $(36, 41)$ and $(140, 122)$, respectively. Both

areas are divided into 9×9 non-overlapping regions. For MDML-DCPs-H3 and MDML-DCPs-C1 to C6, we consistently use a patch size $M \times M$ of 40×40 and the number of regions $J \times J$ within the patch to be 4×4.

The two parameters σ_1 and σ_2 for TT are set to 1.4 and 2.0 on FERET, CAS-PEAL-R1, and FGRC 2.0. Since TT in fact slightly degrades the performance of MDML-DCPs on the LFW View 1 data, the photometric normalization step is omitted for the experiment on LFW. The subspace dimensions of WPCA and PCA for all nine features are set to 600. The dimensions of both the between-individual subspace and within-individual subspace of PLDA are set to 100. There are only two remaining sets of parameters to tune over the four databases: the first is the sampling radii R_{in} and R_{ex} of DCP, and the other is the cost parameter c of the linear SVM [Chang and Lin, 2011]. The optimal values of R_{in} and R_{ex} for FERET, CAS-PEAL-R1, and FGRC 2.0 are $R_{in} = 2$ and $R_{ex} = 3$. c is set at 10^{-4} for FRGC 2.0. For LFW, R_{in}, R_{ex}, and c are set to 4, 6, and 1.0, respectively, based on the results on the View 1 data.

2.5.5.1 Face Identification: FERET

As there is no officially defined training set for FERET, we test MDML-DCPs with the WPCA-based classifiers. The nine WPCA projection matrices are trained on the 1,196 FERET gallery images. For simplicity, the similarity scores computed by the nine classifiers are fused by averaging with equal weights. Performance comparison between MDML-DCPs with the state-of-the-art approaches is shown in Table 2.7. The first two approaches use supervised learning algorithms. The others utilize the unsupervised learning algorithm WPCA. The performance of MDML-DCPs is superior even to those that employ a supervised learning algorithm. In particular, MDML-DCPs achieves the best results on the Fc, Dup1, and Dup2 probe sets. On the Fb set, three images are misclassified by MDML-DCPs, two of which are due to the mislabeling of subjects in the database itself.

2.5.5.2 Face Identification: CAS-PEAL-R1

For experiments on CAS-PEAL-R1, we present MDML-DCPs results with both WPCA and PLDA, with DCP set as the baseline. Both WPCA and PLDA are trained on all 1,200 training images. In both approaches, the similarity scores are fused

Table 2.7: Identification Rates for Different Methods on FERET

	Fb	Fc	Dup1	Dup2
AMFG07 (L,G) + KDCV [Tan and Triggs, 2007]	98.00	98.00	90.00	85.00
TIP10 LGBP + LGXP + LDA [Xie et al., 2010]	99.00	99.00	94.00	93.00
BMVC12 G-LQP + WPCA [Hussain et al., 2012]	**99.90**	**100.0**	93.20	91.00
PAMI14 DFD + WPCA [Lei et al., 2014]	99.40	**100.0**	91.80	92.30
MDML-DCPs + WPCA	99.75	**100.0**	**96.12**	**95.73**

by simple averaging without weighting. A comparison between the performance of MDML-DCPs with other state-of-the-art methods is presented in Table 2.8. We make the following observations:

- "MDML-DCPs + WPCA" achieves the best performance for eight of the nine probe sets of CAS-PEAL-R1. In particular, it outperforms DCP by 32.67% on the challenging PL probe set.
- Compared with DCP, there is limited performance promotion made by "MDML-DCPs + WPCA" on the PU and PD sets. The reason is that the training set contains only frontal images while the PU and PD images possess large pose variations. Therefore the two probe sets benefit little from training.
- The identification rates by "MDML-DCPs + PLDA" are lower than "MDML-DCPs + WPCA", which suggests that there might be a gap between the distribution of training and testing data. In particular, the discriminative subspace learnt by PLDA on the frontal images does not generalize to the three non-frontal probe sets.

2.5.5.3 Face Verification: FRGC 2.0

The performance of MDML-DCPs on Experiments 1 and 4 of the FRGC 2.0 database is discussed. The FRGC 2.0 database provides a large training set, on which PLDA and linear SVM are trained. For each subject in the training set, 12 images are randomly selected and there are 2,664 images in total for SVM training, for which 13,320 intra-personal images pairs and 13,320 inter-personal image pairs are generated according

Table 2.8: Rank-1 Identification Rates for Different Methods on the Nine Probe Sets of PEAL

	PE	PA	PL	PT	PB	PS	PU	PM	PD
TIP07 LGBPHS [Zhang et al., 2007a, 2005]	95.20	86.80	51.00	100.0	98.70	98.90	-	-	-
TIP07 HGPP [Zhang et al., 2007a]	96.80	92.50	62.90	98.40	99.80	99.60	-	-	-
SP09 Weighted LLGP [Xie et al., 2009]	98.00	92.00	55.00	-	-	-	-	-	-
PAMI14 DFD + WPCA [Lei et al., 2014]	99.00	96.90	63.90	-	-	-	-	-	-
DCP + χ^2	96.11	92.82	50.25	100.0	99.10	99.64	65.39	87.44	51.30
MDML-DCPs + PLDA	98.22	97.51	63.26	100.0	**100.0**	100.0	34.31	70.24	33.59
MDML-DCPs + WPCA	**99.62**	**99.21**	**82.92**	**100.0**	99.82	**100.0**	**68.29**	**98.10**	**53.92**

to the method described in [Huang et al., 2007]. All the remaining images are used to train the PLDA model. The performance of the proposed algorithm, together with other state-of-the-art approaches, is shown in Table 2.9. We observe that:

- The MDML-DCPs approach achieves superior performance on both Experiment 1 and 4 with the single descriptor DCP. Score fusion by linear SVM also works slightly better than score averaging.
- On Experiment 1, the performance of MDML-DCPs is nearly perfect. On Experiment 4, it outperforms the current state-of-the-art method [Liu and Liu, 2009] by 0.9%. Moreover, MDML-DCPs utilizes only a single face image descriptor, while the method in [Liu and Liu, 2009] relies on color information and makes use of three descriptors: LBP, Gabor, and Fourier features.

Different from the approaches in Table 2.9, it is shown in [Li et al., 2013b] that by conducting normalization on the similarity matrices which contain similarity scores of each pair of target and query images, verification rates can be improved significantly. However, this operation may not be suitable for the general face verification problems, where only a pair of target and query images is available each time. Therefore, we provide results of MDML-DCPs without score normalization in this chapter.

2.5.5.4 Face Verification: LFW

Experiments in this section follow the image-unrestricted paradigm of the LFW database. Both the PLDA classifier and the more recently proposed Joint Bayesian (JB) classifier [Chen et al., 2012a] are tested for face matching, and both linear SVM and score averaging are tested for the score fusion. For each of the 10-fold cross validations, the remaining nine subsets are partitioned into two: the first eight subsets are used for the training of PCA, PLDA, and JB models, while the last subset is used to train the linear SVM or learn the optimal threshold of similarity scores fused by averaging. According to the description in [Huang et al., 2007], 2,500 intra-personal image pairs and 2,500 inter-personal image pairs are generated using the images in the last subset. No outside training data are employed. In addition, following [Li et al., 2012b, Wolf et al., 2011b], all images with right profile faces are flipped to the left profile faces.

Table 2.9: Verification Rates at 0.1% FAR for Different Methods on the FRGC 2.0 Experiments 1 and 4

		Experiment 1			Experiment 4		
		ROC I	ROC II	ROC III	ROC I	ROC II	ROC III
Single Descriptor	Gabor + LDA [Su et al., 2009]	-	-	97.00	-	-	83.00
	Gabor + Kernel LDA [Tan and Triggs, 2010]	-	-	-	-	-	80.00
	Multiscale LPQ (MLPQ) + Kernel fusion [Chan et al., 2013a]	98.99	98.68	98.37	89.26	89.80	90.36
	Multiscale LBP (MLBP) + Kernel fusion [Chan et al., 2013a]	98.79	98.50	98.21	86.77	87.54	88.21
	MDML-DCPs + PLDA + Score averaging	99.59	99.41	99.22	93.40	93.18	92.89
	MDML-DCPs + PLDA + Linear SVM	**99.61**	**99.43**	**99.25**	**93.91**	**93.64**	**93.39**
Multiple Descriptors	Gabor + LBP + KDCV [Tan and Triggs, 2007]	-	-	-	-	-	83.60
	Gabor + DCT + LDA [Su et al., 2009]	-	-	98.00	-	-	89.00
	Hybrid RCrQ with Gabor+MLBP+DCT [Liu and Liu, 2009]	-	-	-	-	-	92.43
	Gabor + LBP + Kernel LDA [Tan and Triggs, 2010]	-	-	-	-	-	88.10
	LGBP + LGXP + Block-based LDA [Xie et al., 2010]	98.60	98.00	97.30	83.90	84.70	85.20
	Hybrid Fourier feature + LDA [Hwang et al., 2011]	96.38	95.11	93.90	81.82	81.50	81.14
	MLPQ + MLBP + Kernel fusion [Chan et al., 2013a]	99.04	98.77	98.49	90.30	90.94	91.59

A comparison of the classification accuracies is presented in Table 2.10. All the methods in Table 2.10 follow the image-unrestricted training paradigm and do not use outside training data. The ROC curves are plotted in Fig. 2.11. We make the following observations:

- MDML-DCPs outperforms the current state-of-the-art method [Chen et al., 2013] by over 2.2%. It is worth noting that [Chen et al., 2013] also employs dense facial landmarks for face representation. MDML-DCPs outperforms [Li et al., 2012b], which also employs the PLDA-based classifier, by over 5%. The above facts suggest that the MDML-DCPs face representation scheme is more effective than previous approaches.

- Score fusion by linear SVM results in better performance than using score averaging. Moreover, with either linear SVM or score averaging, MDML-DCPs is able to achieve better performance than existing methods, and using both fusion methods the standard error of classification accuracy is consistently smaller than the other approaches.

- The JB-based classifiers slightly outperform PLDA-based classifiers. By fusing the 18 similarity scores produced by the two kinds of classifiers with linear SVM, we achieve a better mean verification rate of 95.58%.

2.6 Conclusion

Due to the degradation of face image quality and large variations of illumination, pose, and expression, the recognition of unconstrained face images is a challenging task. Solving this problem demands work in at least two areas: development of an effective face image descriptor and a comprehensive face representation scheme. To achieve this goal, we make the following contributions:

- We presented a novel face image descriptor named Dual-Cross Patterns. DCP encodes second-order statistics in the most informative directions within a face image. Experimentation on four large-scale face databases shows that DCP has superior discriminative power and is robust to pose, expression, and moderate variations in illumination.

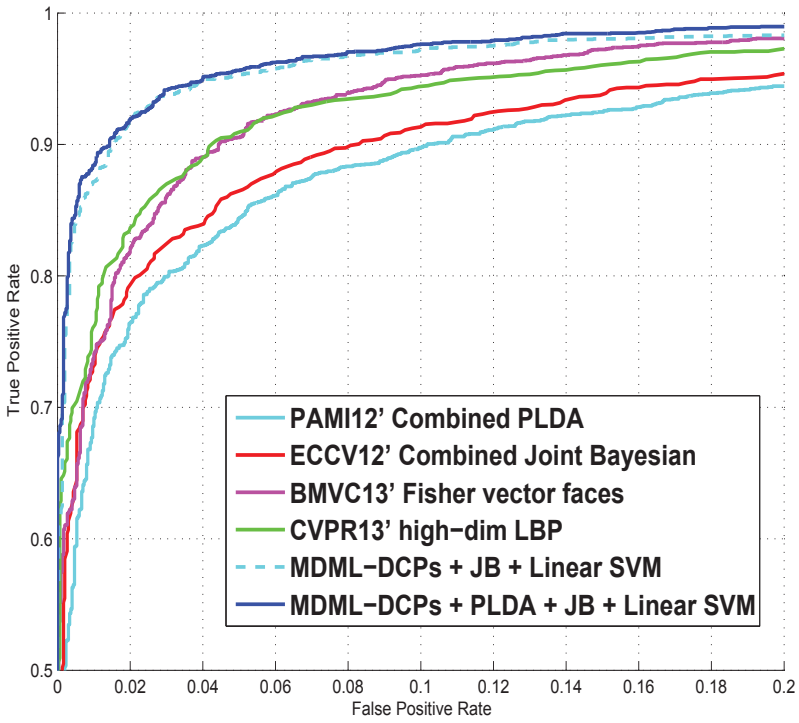


Figure 2.11: ROC curves of the MDML-DCPs method and other state-of-the-art methods in the unrestricted paradigm.

Table 2.10: Mean Verification Accuracy on the LFW View 2 Data

	Accuracy(%) $\pm S_E$
PAMI12 Combined PLDA [Li et al., 2012b]	90.07 $\pm$ 0.51
ECCV12 Combined Joint Bayesian [Chen et al., 2012a]	90.90 $\pm$ 1.48
ICCV13 VMRS [Barkan et al., 2013b]	92.05 $\pm$ 0.45
BMVC13 Fisher vector faces [Simonyan et al., 2013]	93.03 $\pm$ 1.05
CVPR13 high-dim LBP + JB [Chen et al., 2013]	93.18 $\pm$ 1.07
MDML-DCPs + PLDA + Score averaging	**94.57 $\pm$ 0.30**
MDML-DCPs + PLDA + Linear SVM	**95.13 $\pm$ 0.33**
MDML-DCPs + JB + Linear SVM	**95.40 $\pm$ 0.33**
MDML-DCPs + PLDA + JB + Linear SVM	**95.58 $\pm$ 0.34**

- A comprehensive and systematic comparison of fourteen state-of-the-art face image descriptors is conducted on the four face databases. Detailed analysis is provided. Conclusions about the advantages and disadvantages of the descriptors can be drawn from these results.

- The proposed MDML-DCPs face representation scheme comprehensively incorporates both holistic-level DCP features and component-level DCP features. Exploiting the single descriptor DCP, MDML-DCPs consistently achieves the best results on the four databases. In particular, MDML-DCPs improves the verification rate of ROC III in Experiment 4 of FRGC 2.0 to 93.39% and outperforms the state-of-the-art result on LFW by 2.4%.

This work helps to expedite the design of practical face image descriptors and face representation schemes.

Chapter 3

Multimodal Deep Face Representation

Face images appeared in many real-world applications, e.g., social networks and digital entertainment, usually exhibit dramatic pose, illumination, and expression variations, resulting in considerable performance degradation for traditional face recognition algorithms. This chapter proposes a comprehensive deep learning framework to jointly learn face representation using multimodal information. The proposed deep learning structure is composed of a set of elaborately designed convolutional neural networks (CNNs) and a three-layer stacked auto-encoder (SAE). The set of CNNs extracts complementary facial features from multimodal data. Then, the extracted features are concatenated to form a high-dimensional feature vector, whose dimension is compressed by SAE. All the CNNs are trained using a subset of 9,000 subjects from the publicly available CASIA-WebFace database, which ensures the reproducibility of this work. Using the proposed single CNN architecture and limited training data, 98.43% verification rate is achieved on the LFW database. Benefited from the complementary information contained in multimodal data, our small ensemble system achieves higher than 99.0% recognition rate on LFW using publicly available training set.

3.1 Introduction

Recognizing face images captured in real-world circumstances is difficult, because faces usually exhibit rich variations in pose, illumination, expression, and occlusion, as illustrated in Fig. 3.1. In this chapter, inspired by the great success of deep learning

Figure 3.1: Face images in real-world applications usually exhibit rich variations in pose, illumination, expression, and occlusion.

in computer vision, we study the deep learning-based face recognition.

The existing works of deep learning on face recognition extract features from limited modalities, the complementary information contained in more modalities is not well studied. Inspired by the complementary information contained in multi-modalities and the recent progress of deep learning on various fields of computer vision, we present a novel face representation framework that adopts an ensemble of CNNs to leverage the multimodal information. The performance of the proposed multimodal system is optimized from two perspectives. First, the architecture for single CNN is elaborately designed and optimized with extensive experimentations. Second, a set of CNNs is designed to extract complementary information from multiple modalities, i.e., the holistic face image, the rendered frontal face image by 3D model, and uniformly sampled face patches. Besides, we design different structures for different modalities, i.e., a complex structure is designed for the modality that contains the richest information while a simple structure is proposed for the modalities with less information. In this way, we strike a balance between recognition performance and efficiency. The capacity of each modality for face recognition is also compared and discussed.

We term the proposed deep learning-based face representation scheme as Multi-modal Deep Face Representation (MM-DFR), as illustrated in Fig. 3.2. Under this framework, the face representation of one face image involves feature extraction using

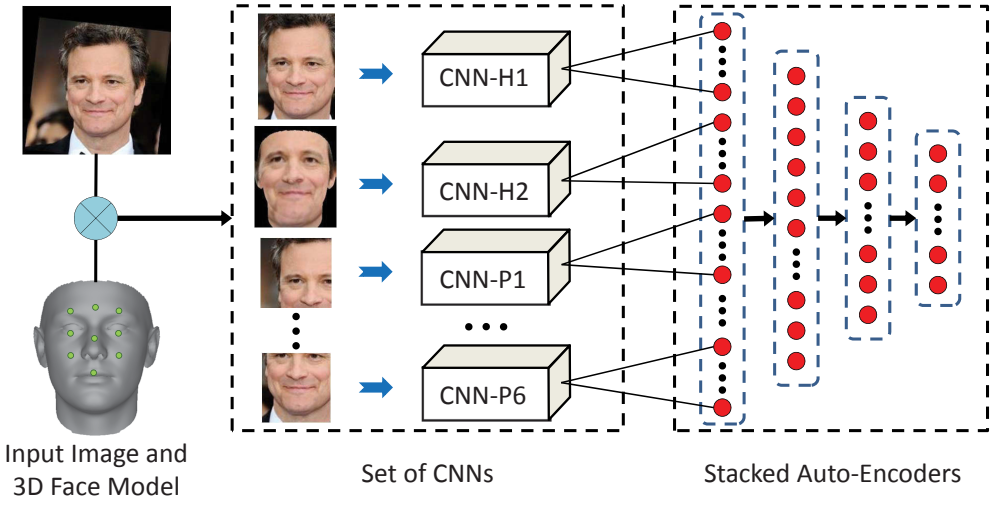


Figure 3.2: Flowchart of the proposed multimodal deep face representation (MM-DFR) framework. MM-DFR is essentially composed of two steps: multimodal feature extraction using a set of CNNs, and feature-level fusion of the set of CNN features using SAE. CNN-H1 is deeper than the other CNNs.

each of the designed CNNs. The extracted features are concatenated as the raw feature vector, whose dimension is compressed by a three-layer SAE. Extensive experiments on the Labeled Face in the Wild (LFW) [Huang et al., 2007] and CASIA-WebFace databases [Yi et al., 2014] indicate that superior performance is achieved with the proposed MM-DFR framework. Besides, the influence of several implementation details, e.g., the usage strategies of ReLU nonlinearity, multiple modalities, aggressive data augmentation, multi-stage training, and L2 normalization, is compared and discussed in the experimentation section. To the best of our knowledge, this is the first published approach that achieves higher than 99.0% recognition rate using a publicly available training set on the LFW database.

The remainder of the chapter is organized as follows: Section 3.2 briefly reviews related works for face recognition and deep learning. The proposed MM-DFR face representation scheme is illustrated in Section 3.3. Face matching using MM-DFR is described in Section 3.4. Experimental results are presented in Section 3.5, leading to conclusions in Section 3.6.

3.2 Related Studies

3.2.1 Face Image Representation

The complicated facial appearance variations call for non-linear techniques for robust face representation, and recent progress on deep learning provides an effective tool. In the following, we review the most relevant progress for deep learning-based face recognition. Taigman et al. [2014a] proposed the DeepFace architecture for face recognition. They used the softmax loss, i.e., the face identification loss, as the supervisory signal to train the network and achieved high recognition accuracy approaching the human-level. Sun et al. [2014] proposed to combine the identification and verification losses for more effective training. They empirically verified that the combined supervisory signal is helpful to promote the discriminative power of extracted CNN features. Zhou et al. [2015] investigated the influence of distribution and size of training data to the performance of CNN. With a huge training set composed of 5 millions of labelled faces, they achieved an accuracy of 99.5% accuracy on LFW using naive CNN structures. One common problem for the above works is that they all employ private face databases for training. Due to the distinct size and unknown distribution of these private data, the performance of the above works may not be directly comparable. Recently, Yi et al. [2014] released the CASIA-WebFace database which contains 494,414 labeled images of 10,575 subjects. The availability of such a large-scale database enables researchers to compete on a fair starting line. In this chapter, the training of all CNNs are conducted exclusively on a subset of 9,000 subjects of the CASIA-WebFace database, which ensures the reproducibility of this work. The CNN architectures designed in this chapter are inspired by two previous works [Simonyan and Zisserman, 2014, Yi et al., 2014], but with a number of modifications and improvements, and our designed CNN models have visible advantage in performance.

3.2.2 Multimodal-based Face Recognition

Most of face recognition algorithms extract a single face representation from the face image. However, they are restrictive in capturing the diverse information contained in the face image. To handle this problem, Ding et al. [2016] proposed to extract

the Multi-directional Multi-level DCPs (MDML-DCPs) feature which includes three holistic-level features and six component-level features. The set of the nine facial features composes the face representation. Similar strategies have been adopted in deep learning-based face representations. For example, the DeepFace approach [Taigman et al., 2014a] adopts the same CNN structure to extract facial features from RGB image, gray-level image and gradient map. The set of face representations are fused in the score level. Sun et al. [2014] proposed to extract deep features from 25 image patches cropped with various scales and positions. The dimension of the concatenated deep features is reduced by Principle Component Analysis (PCA). Multimodal systems that fuse multiple feature cues are also employed in other topics of computer vision, e.g., visual tracking and image classification.

Our multimodal face recognition system is related to the previous approaches, and there is clear novelty. First, we extract multimodal features from the holistic face image, rendered frontal face by 3D face model, and uniformly sampled image patches. The three modalities stand for holistic facial features and local facial features, respectively. Different from [Taigman et al., 2014a] that employs the 3D model to assist 2D piece-wise face warping, we utilize the 3D model to render a frontal face in 3D domain, which indicates much stronger alignment compared with [Taigman et al., 2014a]. Different from [Sun et al., 2014] that randomly crops 25 patches over the face image using dense facial feature points, we uniformly sample a small number of patches with the help of 3D model and sparse facial landmarks, which is more reliable compared with dense landmarks. Second, we propose to employ SAE to compress the high-dimensional deep feature into a compact face signature. Compared with the traditional PCA approach for dimension reduction, SAE has advantage in learning non-linear feature transformations. Third, the large-scale unconstrained face identification problem has not been well studied due to the lack of appropriate face databases. Fortunately, the recently published CASIA-WebFace [Yi et al., 2014] database provides the possibility for such kind of evaluation. In this chapter, we evaluate the identification performance of MM-DFR on the CASIA-WebFace database.

3.3 Multimodal Deep Face Representation

In this section, we describe the proposed MM-DFR framework for face representation. As shown in Fig. 3.2, MM-DFR is essentially composed of two steps: multimodal feature extraction using a set of CNNs, and feature-level fusion of the set of CNN features using SAE. In the following, we describe the two main components in detail.

3.3.1 Single CNN Architecture

All face images employed in this chapter are first normalized to 230×230 pixels with an affine transformation according to the coordinates of five sparse facial feature points, i.e., both eye centers, the nose tip, and both mouth corners. Sample images after the affine transformation are illustrated in Fig. 3.1. We employ an off-the-shelf face alignment tool [Zhang et al., 2014] for facial feature detection. Based on the normalized image, one holistic face image of size 165×120 pixels (Fig. 3.3a) and six image patches of size 100×100 pixels (Fig. 3.3b) are sampled. Another holistic face image is obtained by 3D pose normalization using OpenGL [Ding et al., 2015]. Pose variation is reduced in the rendered frontal face, as shown in Fig. 3.3a.

Two CNN models named NN1 and NN2 are designed, which are closely related to the ones proposed in [Simonyan and Zisserman, 2014, Yi et al., 2014], but with a number of modifications and improvements. We denote the CNN that extracts feature from the holistic face image as CNN-H1. In the following, we take CNN-H1 for example to illustrate the architectures of NN1 and NN2, as shown in Table 3.1 and Table 3.2, respectively. The other seven CNNs employ similar structure but with modifications in resolution for each layer. The major difference between NN1 and NN2 is that NN2 is both deeper and wider than NN1. With larger structure, NN2 is more robust to highly non-linear facial appearance variations; therefore, we apply it to CNN-H1. NN1 is smaller but more efficient and we apply it to the other seven CNNs, with the underlying assumption that the image patches and pose normalized face contain less nonlinear appearance variations. Compared with NN1, NN2 is more vulnerable to overfitting due to its larger number of parameters. In this chapter, we make use of aggressive data augmentation and multi-stage training strategies to reduce overfitting. Details of the two strategies are described in the experimentation section.

NN1 contains 10 convolutional layers, 4 max-pooling layers, 1 mean-pooling layer, and 2 fully-connected layers. In comparison, NN2 incorporates 12 convolutional layers. Small filters of 3×3 are utilized for all convolutional layers. As argued in [Simonyan and Zisserman, 2014], successive convolutions by small filters equal to one convolution operation by a large filter, but effectively enhances the model's discriminative power and reduces the number of filter parameters to learn. ReLU [Dahl et al., 2013] activation function is utilized after all but the last convolutional layers. The removal of ReLU nonlinearity helps to generate dense features, as described in [Yi et al., 2014]. We also remove the ReLU nonlinearity after Fc6; therefore the projection of convolutional features by Fc6 layer is from dense to dense, which means that Fc6 effectively equals to a linear dimension reduction layer that is similar to PCA or Linear Discriminative Analysis (LDA). This is different from previous works that favor sparse features produced by ReLU [Sun et al., 2014, 2015, Taigman et al., 2014a]. Our model is also different from [Yi et al., 2014] since [Yi et al., 2014] simply removes the linear dimension reduction layer (Fc6). The output of the Fc6 layer is employed as face representation. In the experimental section, we empirically justify that the dense-to-dense projection by Fc6 is advantageous to produce more discriminative features. The forward function of ReLU is represented as

$$R(x_i) = max(0, W_c^T x_i + b_c), \tag{3.1}$$

where x_i, W_c, and b_c are the input, weight, and bias of the corresponding convolutional layer before the ReLU activation function. $R(x_i)$ is the output of the ReLU activation function. The dimension of the Fc6 layer is set to 512. The dimension of the Fc7 is set to 9000, which equals to the number of training subjects employed in this chapter. We employ dropout [Krizhevsky et al., 2012] as a regularizer on the first fully-connected layer in the case of overfitting caused by the large amount of parameters. The dropout ratio is set to 0.4. Since this low-dimensional face representation is utilized to distinguish as large as 9,000 subjects in the training set, it should be very discriminative and has good generalization ability.

The other holistic image is rendered by OpenGL with the help of 3D generic face model [Ding et al., 2015]. Pose variation is reduced in the rendered image. We denote the CNN that extracts deep feature from this image as CNN-H2, as illustrated in

Table 3.1: Details of the model architecture for NN1

Name	Type	Input Size	Filter Number	Filter Size /stride /pad	With Relu
Conv11	conv	165×120	64	3×3 /1 /0	yes
Conv12	conv	163×118	128	3×3 /1 /0	yes
Pool1	max pool	161×116	N/A	2×2 /2 /0	no
Conv21	conv	80×58	64	3×3 /1 /0	yes
Conv22	conv	78×56	128	3×3 /1 /0	yes
Pool2	max pool	76×54	N/A	2×2 /2 /0	no
Conv31	conv	38×27	64	3×3 /1 /1	yes
Conv32	conv	38×27	128	3×3 /1 /1	yes
Pool3	max pool	38×27	N/A	2×2 /2 /1	no
Conv41	conv	20×14	128	3×3 /1 /1	yes
Conv42	conv	20×14	256	3×3 /1 /1	yes
Pool4	max pool	20×14	N/A	2×2 /2 /0	no
Conv51	conv	10×7	128	3×3 /1 /1	yes
Conv52	conv	10×7	256	3×3 /1 /1	**no**
Pool5	mean pool	10×7	N/A	2×2 /2 /1	no
Dropout	dropout	6144×1	N/A	N/A	N/A
Fc6	fully-conn	512×1	N/A	N/A	**no**
Fc7	fully-conn	9000×1	N/A	N/A	no
Softmax	softmax	9000×1	N/A	N/A	N/A

Fig. 3.2. Therefore, the first two CNNs encode holistic image features from different modalities. The CNNs that extract features from the six image patches are denoted as CNN-P1, CNN-P2, to CNN-P6, respectively, as illustrated in Fig. 3.2. Exactly the same network structure is adopted for each of the six CNNs.

Different from previous works that randomly sample a large number of image patches [Sun et al., 2014], we propose to sample a small number of image patches uniformly in the semantic meaning, which contributes to maximizing the complementary information contained within the sampled patches. However, the uniform sampling of

Table 3.2: Details of the model architecture for NN2

Name	Type	Input Size	Filter Number	Filter Size /stride /pad	With Relu
Conv11	conv	165×120	64	3×3 /1 /0	yes
Conv12	conv	163×118	128	3×3 /1 /0	yes
Pool1	max pool	161×116	N/A	2×2 /2 /0	no
Conv21	conv	80×58	64	3×3 /1 /0	yes
Conv22	conv	78×56	128	3×3 /1 /0	yes
Pool2	max pool	76×54	N/A	2×2 /2 /0	no
Conv31	conv	38×27	128	3×3 /1 /1	yes
Conv32	conv	38×27	128	3×3 /1 /1	yes
Pool3	max pool	38×27	N/A	2×2 /2 /1	no
Conv41	conv	20×14	256	3×3 /1 /1	yes
Conv42	conv	20×14	256	3×3 /1 /1	yes
Conv43	conv	20×14	256	3×3 /1 /1	yes
Pool4	max pool	20×14	N/A	2×2 /2 /0	no
Conv51	conv	10×7	256	3×3 /1 /1	yes
Conv52	conv	10×7	256	3×3 /1 /1	yes
Conv53	conv	10×7	256	3×3 /1 /1	**no**
Pool5	mean pool	10×7	N/A	2×2 /2 /1	no
Dropout	dropout	6144×1	N/A	N/A	N/A
Fc6	fully-conn	512×1	N/A	N/A	**no**
Fc7	fully-conn	9000×1	N/A	N/A	no
Softmax	softmax	9000×1	N/A	N/A	N/A

the image patches is not easy due to the pose variations of the face appeared in real-world images, as shown in Fig. 3.1. We tackle this problem with a recently proposed strategy for pose-invariant face recognition [Yi et al., 2013]. The principle of the patch sampling process is illustrated in Fig. 3.4. In brief, nine 3D landmarks are manually labeled on a generic 3D face model and the 3D landmarks spread uniformly across the face model. In this chapter, we consistently employ the mean shape of the Basel Face

Figure 3.3: The normalized holistic face images and image patches as input for MM-DFR. (a) The original holistic face image and the 3D pose normalized holistic image; (b) Image patches uniformly sampled from the original face image. Due to facial symmetry and the augmentation by horizontal flipping, we only leverage the six patches illustrated in the first two columns.

Model as the generic 3D face model [Paysan et al., 2009]. Given a 2D face image, it is first aligned to the generic 3D face model using orthogonal projection with the help of five facial feature points. Then, the pre-labeled 3D landmarks are projected to the 2D image. Lastly, a patch of size 100×100 pixels is cropped centering around each of the projected 2D landmarks. More examples of the detected 2D uniform landmarks are shown in Fig. 3.5. It is clear that the patches are indeed uniformly sampled in the semantic meaning regardless of the pose variations of the face image.

3.3.2 Combination of CNNs using Stacked Auto-Encoder

We denote the features extracted by the set of CNNs as $\{x_1, x_2, \cdots, x_K\}$, where $x_i \in \mathbb{R}^{d\times 1}, 1 \leq i \leq K$. In this chapter, K equals to 8 and d equals to 512. The set of features represents multimodal information for face recognition. We conduct feature-level fusion to obtain a single signature for each face image. In detail, the features extracted by the eight CNNs are concatenated as a large feature vector, denoted as:

$$\hat{x} = [x_1; x_2; \cdots; x_K] \in \mathbb{R}^{Kd\times 1}. \tag{3.2}$$

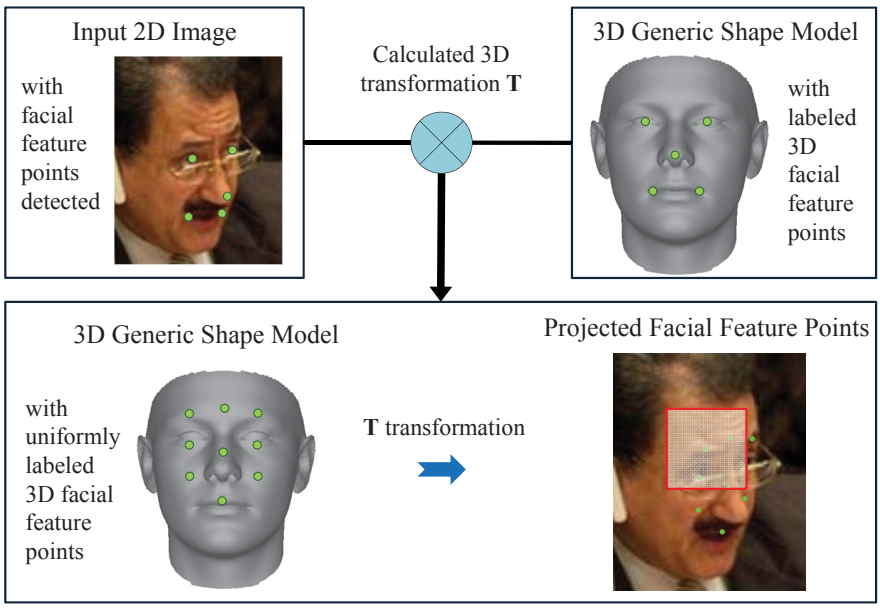


Figure 3.4: The principle of patch sampling adopted in this chapter. A set of 3D landmarks are uniformly labeled on the 3D face model, and are projected to the 2D image. Centering around each landmark, a square patch of size 100×100 pixels is cropped, as illustrated in Fig. 3.3b.

Figure 3.5: More examples about the uniformly detected landmarks that are projected from a generic 3D face model to 2D images.

$\hat{x}$ is high dimensional, which is impractical for real-world face recognition applications. We further propose to reduce the dimension of $\hat{x}$ by SAE. Compared with the traditional dimension reduction approaches, e.g., PCA, SAE has advantage in learning non-linear feature transformations. In this chapter, we employ a three-layer SAE. The number of the neurons of the three auto-encoders are 2048, 1024, and 512, respectively. The output of the last encoder is utilized as the compact signature of the face image. The structure for the designed SAE is illustrated in Fig. 3.2.

Nonlinear activation function is utilized after each of the fully-connected layers. Two activation functions, i.e., sigmoid function and hyperbolic tangent (tanh) function, are evaluated. The forward function of the sigmoid activation function is represented as

$$S(x_i) = \frac{1}{1+exp(-W_f^T x_i - b_f)}. \tag{3.3}$$

The forward function of the tanh activation function is represented as

$$T(x_i) = \frac{exp(W_f^T x_i + b_f) - exp(-W_f^T x_i - b_f)}{exp(W_f^T x_i + b_f) + exp(-W_f^T x_i - b_f)}, \tag{3.4}$$

where x_i, W_f, and b_f are the input, weight, and bias of the corresponding fully-connected layer before the activation function. Different normalization schemes of $\hat{x}$ are adopted for the sigmoid and tanh activation functions, since their output space is different. For the sigmoid function, we normalize the elements of $\hat{x}$ to be within $[0, 1]$. For the tanh function, we normalize the elements of $\hat{x}$ to be within $[-1, +1]$. In the experimentation section, we empirically compare the performance of SAE with the two different nonlinearities.

3.4 Face Matching with MM-DFR

In this section, the face matching problem is addressed based on the proposed MM-DFR framework. Two evaluation modes are adopted: the unsupervised mode and the supervised mode. Suppose two features produced by MM-DFR for two images are denoted as y_1 and y_2, respectively. In the unsupervised mode, the cosine distance is

employed to measure the similarity s between y_1 and y_2.

$$s(y_1, y_2) = \frac{y_1^T y_2}{\|y_1\|\|y_2\|}. \tag{3.5}$$

For the supervised mode, a number of discriminative or generative models can be employed [Chen et al., 2012a, Prince and Elder, 2007], In this chapter, we employ the Joint Bayesian (JB) model [Chen et al., 2012a] as it is shown to outperform other popular models in recent works [Ding et al., 2016]. For both the unsupervised and supervised modes, the nearest neighbor (NN) classifier is adopted for face identification. JB models the face generation process as

$$x = \mu + \varepsilon, \tag{3.6}$$

where μ represents the identity of the subject, while ε represents intra-personal noise.

JB solves the face identification or verification problems by computing the log-likelihood ratio between the probability $P(x_1, x_2|H_I)$ that two faces belong to the same subject and the probability $P(x_1, x_2|H_E)$ that two faces belong to different subjects.

$$r(x_1, x_2) = log\frac{P(x_1, x_2|H_I)}{P(x_1, x_2|H_E)}, \tag{3.7}$$

where $r(x_1, x_2)$ represents the log-likelihood ratio, and we refer to $r(x_1, x_2)$ as similarity score for clarity in the experimental part of the chapter.

3.5 Experiments

In this section, extensive experiments are conducted to present the effectiveness of the proposed MM-DFR framework. The experiments are conducted on two large-scale unconstrained face databases, i.e., LFW [Huang et al., 2007] and CASIA-WebFace [Yi et al., 2014]. Images in both databases are collected from internet; therefore they are images that appear in real-world circumstances.

The LFW [Huang et al., 2007] database contains 13,233 images of 5,749 subjects. Images in this database exhibit rich intra-personal variations of pose, illumination, and expression. It has been extensively studied for the research of unconstrained

face recognition in recent years. Images in LFW are organized into two "Views". View 1 is for model selection and parameter tuning while View 2 is for performance reporting. In this chapter, we follow the official protocol of LFW and report the mean verification accuracy and the standard error of the mean (S_E) by the 10-fold cross-validation scheme on the View 2 data.

Despite of its popularity, the LFW database contains limited number of images and subjects, which restricts its evaluation for large-scale unconstrained face recognition applications. The CASIA-WebFace [Yi et al., 2014] database has been released recently. CASIA-WebFace contains 494,414 images of 10,575 subjects. As images in this database are collected in a semi-automatic way, there is a small amount of mis-labeled images in this database. Because there is no officially defined protocol for face recognition on this database, we define our own protocol for face identification in this chapter. In brief, we divide CASIA-WebFace into two sets: a training set and a testing set. The 10,575 subjects are ranked in the descent order by the number of their images contained in the database. The 471,592 images of the top 9,000 subjects compose the training set. The 22,822 images of the rest 1,575 subjects make up the testing set.

All CNNs and SAE in this chapter are trained using the 9,000 subjects in the defined training set above. Images are converted to gray-scale and geometrically normalized as described in Section 3.3. For NN1, we double the size of the training set by flipping all training images horizontally to reduce overfitting. Therefore, the size of training data for NN1 is 943,184. For NN2, we adopt much more aggressive data augmentation by horizontal flipping, image jittering,[1], and image down-sampling. The size of the augmented training data for NN2 is about 1.8 million. The distribution of training data for NN1 and NN2 is illustrated in Fig. 3.6. It is shown that the long-tail distribution characteristic [Zhou et al., 2015] of the original training data is improved after the aggressive data augmentation for NN2.

We adopt the following multi-stage training strategy to train all the CNN models. First, we train the CNN models as a multi-class classification problem, i.e., softmax loss is employed. For all CNNs, the initial learning rate for all learning layers is set to 0.01, and is divided by 10 after 10 epochs, to the final rate of 0.001. Second, we adopt the recently proposed triplet loss [Schroff et al., 2015] for fine-tuning for 2 more

[1]For image jittering, we add random gaussian noise on the coordinates of the five facial feature points. The noise is distributed with zero mean and standard deviation of four pixels.

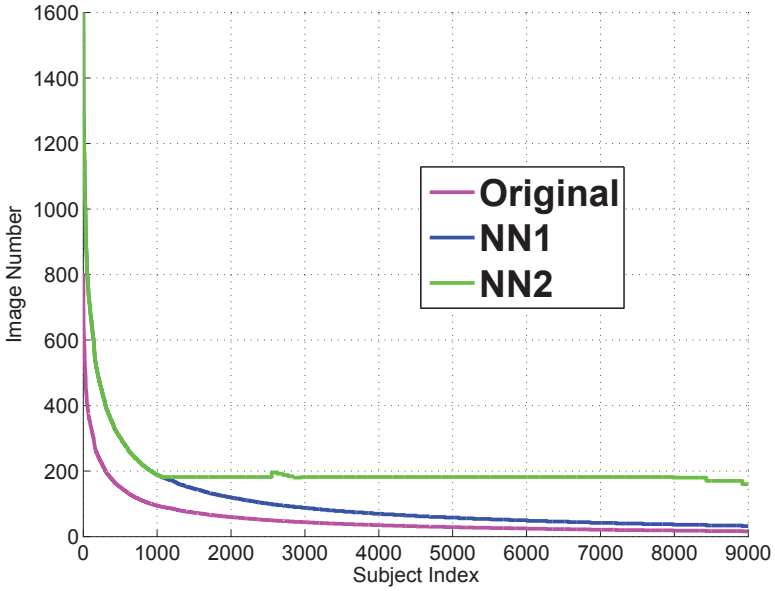


Figure 3.6: Training data distribution for NN1 and NN2. This figure plots the number of images for each subject in the training set. The long-tail distribution characteristic [Zhou et al., 2015] of the original training data is improved after the aggressive data augmentation for NN2.

epochs. We set the margin for the triplet loss to be 0.2 and learning rate to be 0.001. It is expected that this multi-stage training strategy can boost performance while converge faster than using the triplet loss alone [Schroff et al., 2015]. For SAE, the learning rate decreases from 0.01 to 0.00001, gradually. We train each of the three auto-encoders one by one and each auto-encoder is trained for 10 epochs. In the testing phase, we extract deep feature from both the original image and its horizontally flipped image. Unless otherwise specified, the two feature vectors are averaged as the representation of the input face image. The open-source deep learning toolkit Caffe [Jia et al., 2014] is utilized to train all the deep models.

Five sets of experiments are conducted. First, we empirically justify the advantage of dense features for face recognition by excluding two ReLU nonlinearities compared with previous works. The performance of the proposed single CNN model is also compared against the state-of-the-art CNN models on the LFW database. Next, the performance of the eight CNNs contained within the MM-DFR framework is compared on face verification task on LFW. Then, the fusion of the eight CNNs by SAE is conducted and different nonlinearities are also compared. We also test

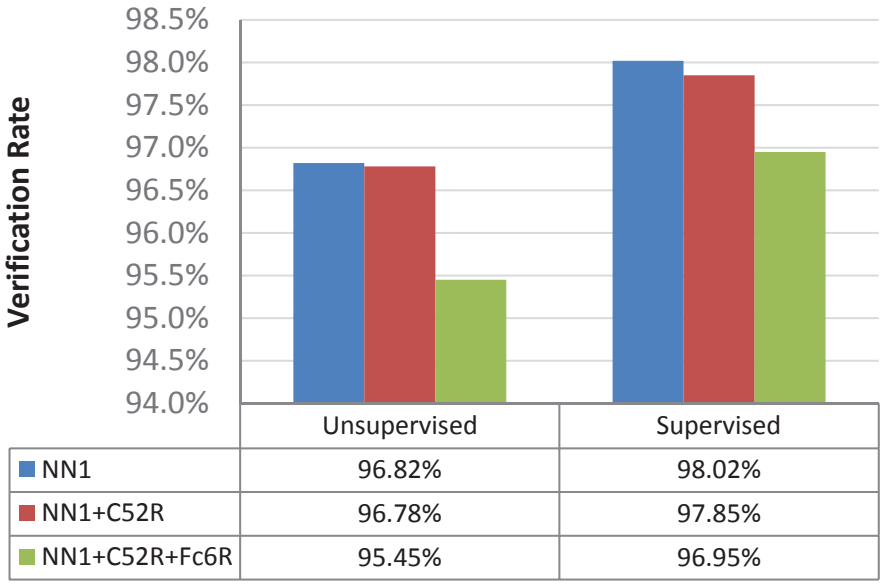
the performance of MM-DFR followed with the supervised classifier JB. Lastly, face identification experiment is conducted on the CASIA-WebFace database with our own defined evaluation protocol.

3.5.1 Performance Comparison with Single CNN Model

In this experiment, we evaluate the role of ReLU nonlinearity using CNN-H1 as an example. For fast evaluation, the comparison is conducted with the simple NN1 structure described in Table 3.1 and only the softmax loss is employed for model training. Performance of CNN-H1 using the NN2 structure can be found in Table 3.4. Two paradigms[1] are followed: 1) the unsupervised paradigm that directly calculate the similarity between two CNN features using cosine distance metric. 2) the supervised paradigm that uses JB to calculate the similarity between two CNN features. For the supervised paradigm, we concatenate the CNN features of the original face image and its horizontally flipped version as the raw representation of each test sample. Then, we adopt PCA for dimension reduction and JB for similarity calculation. The dimension of the PCA subspace is tuned on the View 1 data of LFW and applied to the View 2 data. Both PCA and JB are trained on the CASIA-WebFace database. For PCA, to boost performance, we also re-evaluate the mean of CNN features using the 9 training folds of LFW in 10-fold cross validation.

The performance of three structures are reported in Fig. 3.7 and Fig. 3.8: 1) NN1, 2) NN1 with ReLU after Conv52 layer (denoted as NN1+C52R), and 3) NN1 with ReLU after both Conv52 and Fc6 (denoted as NN1+C52R+Fc6R). For both NN1+C52R and NN1+C52R+Fc6R, we replace the average pooling layer after Conv 52 with max pooling accordingly. It is shown in Fig. 3.7 that the ReLU nonlinearity after Conv52 or Fc6 actually harms the performance of CNN. The experimental results have two implications: 1) dense feature is preferable than sparse feature for CNN, as intuitively advocated in [Yi et al., 2014]. However, there is no experimental justification in [Yi et al., 2014]. 2) the linear projection from the output of the ultimate convolutional layer (Conv52) to the low-dimensional subspace (Fc6) is better than the commonly adopted non-linear projection. This is clear evidence that the negative response of the ultimate

[1] Similar to previous works [Taigman et al., 2014a, Yi et al., 2014], both the two paradigms defined in this chapter correspond to the "Unrestricted, Labeled Outside Data Results" protocol that is officially defined in [Huang et al., 2007].



	Unsupervised	Supervised
NN1	96.82%	98.02%
NN1+C52R	96.78%	97.85%
NN1+C52R+Fc6R	95.45%	96.95%

Figure 3.7: Performance comparison on LFW with different usage strategies of ReLU nonlinearity.

convolutional layer (Conv52) also contains useful information.

The performance by single CNN models on LFW is reported in Table. 3.3. The performance of the state-of-the-art CNN models is also tabulated. Compared with Fig. 3.7, we further improve the performance of NN1 by fine-tuning with triplet loss. It seems that the triplet loss mainly improves the performance for the unsupervised mode in our experiment. It is shown that the proposed CNN model consistently outperforms the state-of-the-art CNN models under both the unsupervised paradigm and supervised paradigm. In particular, compared with [Wang et al., 2015, Yi et al., 2014] that both employ the complete CASIA-WebFace database for CNN training, we only leverage a subset of the CASIA-WebFace database. With more training data, we expect the proposed CNN model can outperform the other models with an even larger margin.

3.5.2 Performance of the Eight CNNs in MM-DFR

In this experiment, we present in Table 3.4 the performance achieved by each of the eight CNNs contained within the MM-DFR framework. We report the performance of CNN-H1 with the NN2 structure while the other seven CNNs all employ the more efficient NN1 structure. The same as the previous experiment, both the unsupervised

Table 3.3: Performance Comparison on LFW using Single CNN Model on Holistic Face Image

	Accuracy (Unsupervised)	Accuracy (Supervised)
DeepFace [Taigman et al., 2014a]	95.92 ± 0.29	97.00 ± 0.87
DeepID2 [Sun et al., 2014]	-	$96.33 \pm$ -
Arxiv2014 [Yi et al., 2014]	96.13 ± 0.30	97.73 ± 0.31
Facebook [Taigman et al., 2014b]	-	$98.00 \pm$ -
MSU TR [Wang et al., 2015]	96.95 ± 1.02	97.45 ± 0.99
Ours (NN1)	$\mathbf{97.32 \pm 0.34}$	$\mathbf{98.05 \pm 0.22}$
Ours (NN2)	$\mathbf{98.12 \pm 0.24}$	$\mathbf{98.43 \pm 0.20}$

Table 3.4: Performance Comparison on LFW of Eight Individual CNNs

	Accuracy (Unsupervised)	Accuracy (Supervised)
CNN-H1	$\mathbf{98.12 \pm 0.24}$	$\mathbf{98.43 \pm 0.20}$
CNN-H2	96.47 ± 0.44	97.67 ± 0.28
CNN-P1	96.83 ± 0.26	97.30 ± 0.22
CNN-P2	97.25 ± 0.31	98.00 ± 0.24
CNN-P3	96.70 ± 0.25	97.82 ± 0.16
CNN-P4	96.17 ± 0.31	96.93 ± 0.21
CNN-P5	96.05 ± 0.27	97.23 ± 0.20
CNN-P6	95.58 ± 0.17	96.72 ± 0.21

paradigm and supervised paradigm are followed. For the supervised paradigm, the PCA subspace dimension of the eight CNNs is unified to be 110. Besides, features of the original face image and the horizontally flipped version are L2 normalized before concatenation. We find that this normalization operation typically boosts the performance of the supervised paradigm by 0.1% to 0.4%.

When combining Table 3.3 and Table 3.4, it is clear that CNN-H1 outperforms CNN-H2 with the same NN1 structure, although they both extract features from holistic face images. This maybe counter-intuitive, since the impact of pose variation has been reduced for CNN-H2. We explain this phenomenon from the following

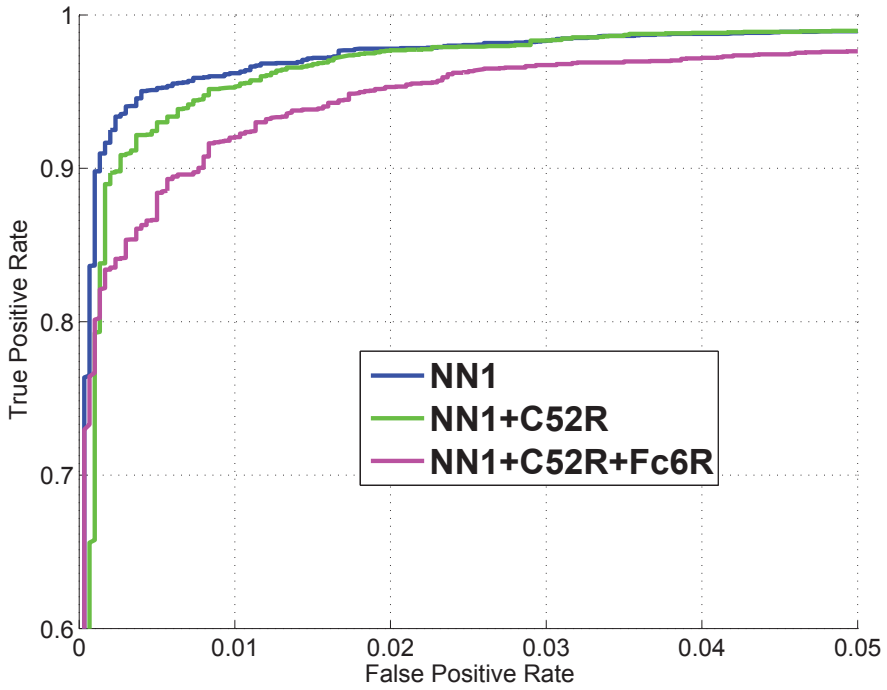


Figure 3.8: ROC curves of different usage strategies of the ReLU nonlinearity on LFW.

two aspects: 1) most images in LFW are near-frontal face images, so the 3D pose normalization employed by CNN-H2 does not contribute much to pose correction. 2) the errors in pose normalization bring about undesirable distortions and artifacts to facial texture, e.g., the distorted eyes, nose, and mouth shown in Fig. 3.3(a). The distorted facial texture is adverse to face recognition, as argued in our previous work [Ding and Tao, 2016]. However, we empirically observe that the performance of MM-DFR drops slightly on View 1 data if we exclude CNN-H2, which indicates CNN-H2 provides complementary information to CNN-H1 from a novel modality. The contribution of CNN-H2 to MM-DFR is also justified by the last experiment in this section. Besides, the performance of the patch-level CNNs, i.e., CNN-P1 to CNN-P6, fluctuates according to the discriminative power of the corresponding patches.

3.5.3 Fusion of CNNs with SAE

In this experiment, we empirically choose the best nonlinearity for SAE that is employed for feature-level fusion of the eight CNNs. The structure of SAE employed

in this chapter is described in Fig. 3.2. For each CNN, we average the features of the original image and the horizontally flipped version. L2 normalization is conducted for each averaged feature before concatenating the features produced by the eight CNNs. Similar to the previous experiment, we find this normalization operation promotes the performance of MM-DFR. The dimension of the input for SAE is 4,096. Two types of non-linearities are evaluated, the sigmoid non-linearity and the tanh non-linearity, denoted as SAE-SIG and SAE-TANH, respectively. The output of the third encoder (before the nonlinear layer) is utilized as the signature of the face image. Cosine distance is employed to evaluate the similarity between two face images. SAE are trained on the training set of CASIA-WebFace, using feature vectors extracted from both the original images and the horizontally flipped ones. The performance of SAE-SIG and SAE-TANH is 98.33% and 97.90% on the View1 data of LFW, respectively.

SAE-TANH considerably outperforms SAE-SIG. One important difference between the sigmoid non-linearity and the tanh non-linearity is that they normalize the elements of the feature to be within $[0, 1]$ and $[-1, 1]$, respectively. Compared with the tanh non-linearity, the sigmoid non-linearity loses the sign information of feature elements. However, the sign information is valuable for discriminative power.

3.5.4 Performance of MM-DFR with Joint Bayesian

The above three experiments have justified the advantage of the proposed CNN structures. In this experiment, we further promote the performance of the proposed framework.

We show the performance of MM-DFR with JB, where the output of MM-DFR is utilized as the signature of the face image. We term this face recognition pipeline as MM-DFR-JB. For comparison, the performance achieved by CNN-H1 with the JB classifier is also presented, denoted as "CNN-H1 + JB". The performance of the two systems is tabulated in Table 3.5 and the ROC curves are illustrated in Fig. 3.9. It is shown that MM-DFR considerably outperforms the single modal-based approach, which indicates the fusion of multimodal information is important to promote the performance of face recognition systems. By excluding the five labeling errors in LFW, the actual performance of MM-DFR-JB reaches 99.10%.

Our simple 8-net based ensemble system also outperforms DeepID2 [Sun et al.,

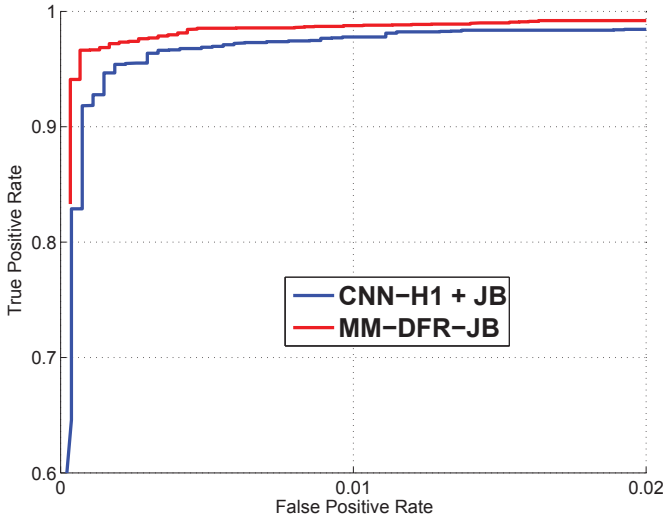


Figure 3.9: Performance comparison between the proposed MM-DFR approach and single modality-based CNN on the face verification task.

Table 3.5: Performance Evaluation of MM-DFR with JB

	#Nets	Accuracy(%)$\pm S_E$
CNN-H1 + JB	1	98.43 $\pm$ 0.20
DeepFace [Taigman et al., 2014a]	7	97.25 $\pm$ 0.81
MSU TR [Wang et al., 2015]	7	98.23 $\pm$ 0.68
DeepID2 [Sun et al., 2014]	25	98.97 $\pm$ 0.25
MM-DFR-JB	8	**99.02 $\pm$ 0.19**

2014], which includes as much as 25 CNNs. Some more recent approaches, e.g. [Schroff et al., 2015, Sun et al., 2015], achieve better performance than MM-DFR. However, they either employ significantly larger private training dataset or considerably larger number of CNN models. In comparison, we employ only 8 nets and train the models using a relatively small training set.

3.5.5 Face Identification on CASIA-WebFace Database

The face identification experiment is conducted on the test data of the CASIA-WebFace database, which includes 22,822 images of 1,575 subjects. For each subject, the first five images are selected to make up the gallery set, All the other images compose the

Table 3.6: The rank-1 identification rates by Different Combinations of Modalities on CASIA-WebFace Database

	Identification Rates
CNN-H1 + JB	72.26%
CNN-H2 + JB	69.07%
CNN-H1&H2 + JB	74.51%
CNN-P1 to P6 + JB	76.01%
MM-DFR-JB	76.53%

probe set. Therefore, there are 7,875 gallery images and 14,947 probe images in total.

The rank-1 identification rates by different combinations of modalities are tabulated in Table 3.6. The corresponding Cumulative Match Score (CMS) curves are illustrated in Fig. 3.10. It is shown that although very high face verification rate has been achieved on the LFW database, large-scale face identification in real-world applications is still a very hard problem. In particular, the rank-1 identification rate by the proposed approach is only 76.53%.

It is clear that the proposed multimodal face recognition algorithm significantly outperforms the single modal based approach. In particular, the rank-1 identification rate of MM-DFR-JB is higher than that of "CNN-H1 + JB" by as much as 4.27%. "CNN-H1 + JB" outperforms "CNN-H2 + JB" with a large margin, partially because CNN-H1 is based on the larger architecture NN2 and trained with more aggressively augmented data. However, the combination of the two modalities still considerably boosts the performance by 2.25% on the basis of CNN-H1, which forcefully justifies the contribution of the new modality introduced by 3D pose normalization. These experimental results are consistent with those obversed on the LFW database. Experimental results on both datasets strongly justify the effectiveness of the proposed MM-DFR framework.

3.6 Conclusion

Face recognition in real-world applications is a challenging task because of the rich appearance change caused by pose, expression, and illumination variations. We handle

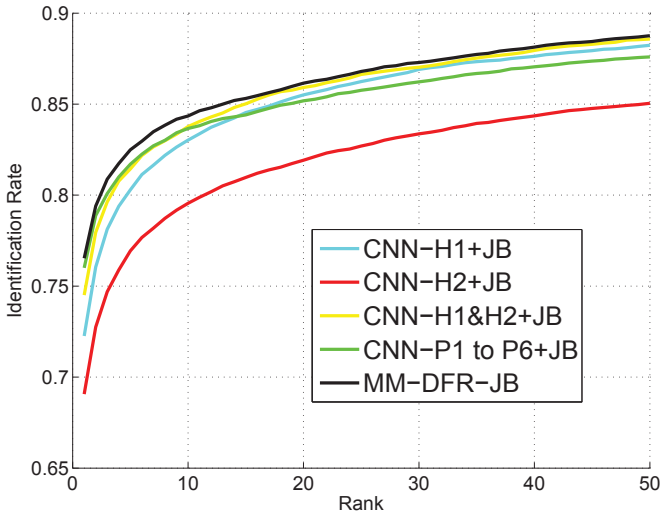


Figure 3.10: CMS curves by different combinations of modalities on the face identification task.

this problem by elaborately designing a deep architecture that employs complementary information from multimodal image data. First, we enhance the recognition ability of each CNN by carefully integrating a number of published or our own developed tricks, such as deep structures, small filters, careful use of ReLU nonlinearity, aggressive data augmentation, dropout, multi-stage training with multiple losses, and L2 normalization. Second, we propose to extract multimodal information using a set of CNNs from the original holistic face image, the rendered frontal pose image by 3D model, and uniformly sampled image patches. Third, we present the feature-level fusion approach using stacked auto-encoders to fuse the features extracted by the set of CNNs, which is advantageous to learn non-linear dimension reduction. Extensive experiments have been conducted for both face verification and face identification experiments. As the proposed MM-DFR approach effectively employs multimodal information for face recognition, clear advantage of MM-DFR is shown compared with the single modal-based algorithms and some state-of-the-art deep models. Other deep learning based approaches may also benefit from the structures that have been proved to be useful in this chapter. In the future, we will try to integrate more multimodal information into the MM-DFR framework and further promote the performance of single deep architecture such as NN2.

Chapter 4

Multi-task Pose-Invariant Face Recognition

Face images captured in unconstrained environments usually contain significant pose variation, which dramatically degrades the performance of algorithms designed to recognize frontal faces. This chapter proposes a novel face identification framework capable of handling the full range of pose variations within $\pm 90^{\circ}$ of yaw. The proposed framework first transforms the original pose-invariant face recognition problem into a partial frontal face recognition problem. A robust patch-based face representation scheme is then developed to represent the synthesized partial frontal faces. For each patch, a transformation dictionary is learnt under the proposed multi-task learning (MTL) scheme, which helps to utilize the correlation between poses. The transformation dictionary transforms the features of different poses into a discriminative subspace. Lastly, face matching is performed at patch level rather than at the holistic level. Extensive and systematic experimentation on FERET, CMU-PIE, and Multi-PIE databases shows that the proposed method consistently outperforms single-task based baselines as well as state-of-the-art methods for the pose problem. We further extend the proposed algorithm for the unconstrained face verification problem and achieve top level performance on the challenging LFW dataset.

4.1 Introduction

Face recognition has been one of the most active research topics in computer vision for more than three decades. With years of effort, promising results have been achieved for automatic face recognition, in both controlled [Tan and Triggs, 2010] and uncontrolled environments [Ding et al., 2016]. However, face recognition remains significantly affected by the wide variations of pose, illumination, and expression often encountered in real-world images. The pose problem in particular is still largely unsolved, as argued in a recent work [Abiantun et al., 2014]. In this chapter, we mainly handle the identification problem of matching an arbitrary pose probe face with frontal gallery faces, which is the most common setting for both the research and application of pose-invariant face recognition (PIFR) [Abiantun et al., 2014, Arashloo and Kittler, 2011, Blanz and Vetter, 2003, Ho and Chellappa, 2013, Prince et al., 2008]. At the end of the chapter, we briefly extend the proposed approach to solve the unconstrained face verification problem [Huang et al., 2007].

Pose variation induces dramatic appearance change in the face image. Essentially, this is caused by the complex 3D geometrical structure of the human head. As shown in Fig. 4.1, the rigid rotation of the head results in self-occlusion, which means that some facial texture will be invisible with variations in pose. Even the shape and position in the image of visible facial texture vary nonlinearly from pose to pose. The pose problem is also usually combined with other factors, such as variations in illumination and expression, to affect the appearance of face images. In consequence, the extent of appearance change caused by pose variation is usually greater than that caused by differences in identity, and the performance of frontal face recognition algorithms degrades dramatically when the images to be matched feature different poses.

Directly matching faces in different poses is difficult. One intuitive solution is to conduct face synthesis so that the two facial images can be compared in the same pose. Most approaches following this idea are dedicated to recovering complete frontal faces from non-frontal faces [Abiantun et al., 2014, Blanz et al., 2005, Chai et al., 2007, Ho and Chellappa, 2013, Zhu et al., 2013]. However, synthesizing the entire frontal face from profile faces is difficult since most facial texture is invisible as a result of occlusion. Therefore, the aforementioned methods tend to constrain their recognition ability within $\pm 45^{\circ}$ of yaw variation.

Figure 4.1: (a) The rigid rotation of the head results in self-occlusion as well as nonlinear facial texture deformation. (b) The pose problem is combined with other factors, e.g., variations in expression and illumination, to affect face recognition.

Inspired by the observation that human beings can easily recognize profile faces without the need to elaborately recover the whole frontal face, we present a novel face representation approach that makes full use of just the facial texture that is occlusion-free. This representation approach is general in nature, which means that it applies continuously to the full range of pose variations between $-90°$ and $+90°$ of yaw. Using simple pose normalization and pre-processing operations, our approach converts the original PIFR problem into a partial face recognition task [Liao et al., 2013], in which the face is represented by the unoccluded facial parts in a patch-based fashion. The proposed face representation scheme is therefore named Patch-based Partial Representation (PBPR).

Feature transformation enhances recognition ability by transforming the features from the gallery and probe images to a common discriminative subspace. Based on this intuition, we propose a learning method called Multi-Task Feature Transformation Learning (MtFTL). By considering the correlation between the transformation matrices for different poses, MtFTL consistently achieves better performance than its single-task based counterparts. Its advantage is particularly evident when the size of training data is limited. The transformation matrices learnt by MtFTL are highly compact as a result of sharing most projection vectors across poses, which additionally reduces memory cost.

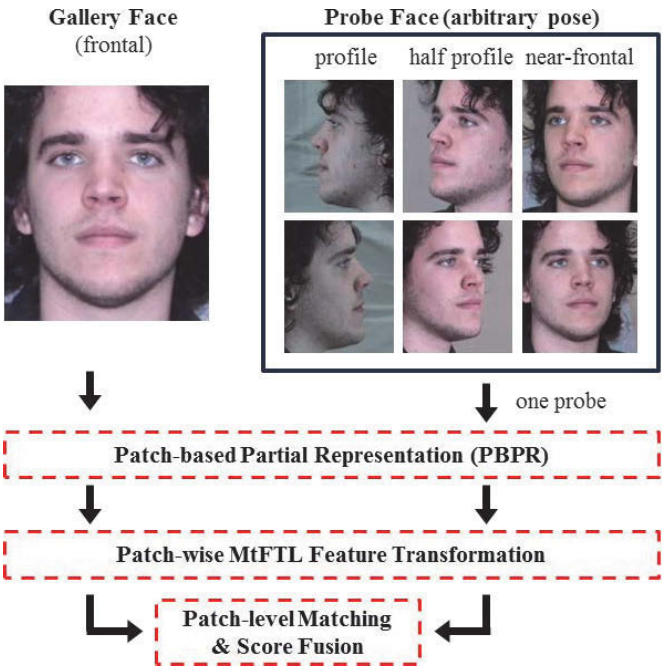


Figure 4.2: Overview of the proposed PBPR-MtFTL framework for pose-invariant face recognition, as applied to the recognition of arbitrary pose probe faces.

We term the entire proposed framework for tackling the pose problem PBPR-MtFTL. Under this framework, matching an arbitrary pose probe face and frontal gallery faces involves transformation of the extracted PBPR representation using the learnt MtFTL transformation dictionaries, followed by patch-level cosine distance computation and score fusion, as illustrated in Fig. 4.2. Extensive experiments on FERET, CMU-PIE, and Multi-PIE datasets indicate that superior performance is consistently achieved with PBPR-MtFTL.

The remainder of the chapter is organized as follows: Section 4.2 briefly reviews related works for multi-task learning. The proposed PBPR face representation scheme is illustrated in Section 4.3. The multi-task feature transformation learning approach MtFTL is described in Section 4.4. Face matching using PBPR-MtFTL is introduced in Section 4.5. Experimental results are presented in Section 4.6, leading to conclusions in Section 4.7.

4.2 Related Studies

Many promising approaches have been proposed to tackle the pose challenge in face recognition. For a comprehensive survey on PIFR methods, please refer to [Ding and Tao, 2016] and Section 1.4 of this thesis. In the following, we only review related multi-task learning approaches.

Multi-task learning (MTL) is a machine learning technique that learns several tasks simultaneously for better performance by capturing the intrinsic correlation between different tasks. MTL implicitly increases the sample size and improves the generalization ability for each task; hence, it is especially beneficial when the training data for the tasks is small.

While MTL has been widely applied to computer vision tasks, e.g., visual tracking [Hong et al., 2013], action recognition [Mahasseni and Todorovic, 2013], and face recognition [Masip et al., 2007, 2008], it is new for PIFR. Existing approaches for PIFR ignore the correlation between the feature transformations of different poses [Li et al., 2009, Sharma et al., 2012]. To the best of our knowledge, MTL for PIFR is only briefly mentioned in [Zhu et al., 2014b] but no detailed information is provided, and multi-view reconstruction is targeted rather than feature transformation learning. Nevertheless, MTL provides a principled way for us to model the correlation between poses if we view the learning of feature transformation for each pose as a task. MtFTL is arguably the first MTL approach that jointly learns feature transformations for different poses and is shown to profit from the latent inter-pose correlations.

4.3 Face Representation for the Pose Problem

Existing face representation methods tend to extract fixed-length features from face images, with the underlying assumption that all facial components are visible in the image [Liao et al., 2013]. However, as shown in Fig. 4.2, this hypothesis does not hold for a profile face where there is severe self-occlusion. In this section, we propose the flexible PBPR face representation scheme, where the length of face representation is related to the pose of the face; for example, a frontal face image will have larger face representation than a profile face image. This is reasonable, since the profile face provides less information for recognition. As shown in Fig. 4.3, PBPR is

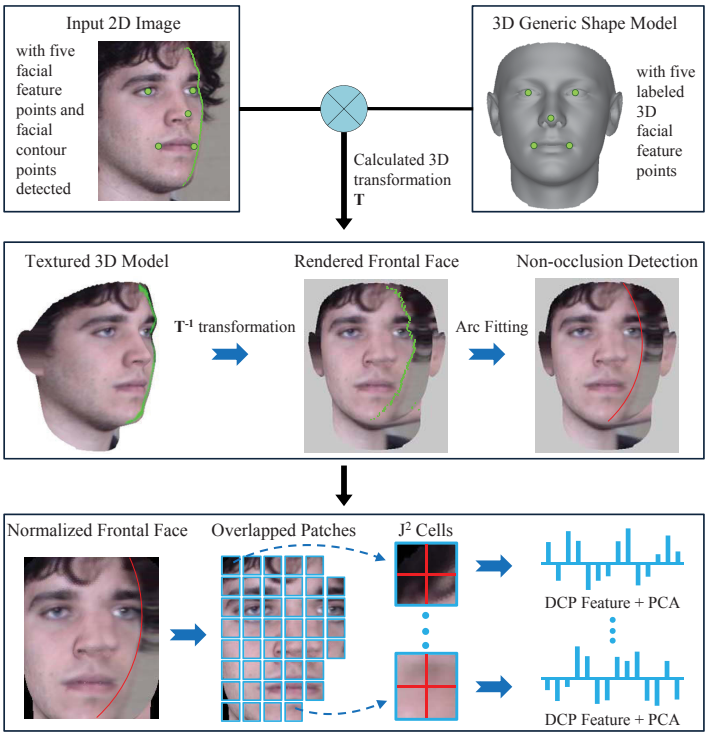


Figure 4.3: Overview of the proposed PBPR face representation method. PBPR is applied to arbitrary pose face images. The final PBPR representation is a set of patch-level DCP features after dimension reduction by PCA.

essentially composed of three steps: face pose normalization, unoccluded facial texture detection, and patch-wise feature extraction. In this section, we describe the three main components in detail.

4.3.1 Face Pose Normalization

A standard 3D method is adopted for face pose normalization [Asthana et al., 2011]. The five most stable facial feature points, i.e., the centers of both eyes, the tip of the nose, and the two mouth corners, are first detected automatically or manually. For profile faces (as shown in Fig. 4.2), the coordinates of the occluded facial feature points are estimated. Using the orthographic projection model [Sun et al., 2013b] and the detected five facial feature points, a 3D generic shape model is aligned to the 2D face image[1]. The 2D face image is then back-projected to the 3D model, and a frontal face image is rendered with the textured 3D model.

Previous works rely on dense facial feature points, e.g., 68 points in [Asthana et al., 2011] and 79 points in [Abiantun et al., 2014], for accurate pose normalization. However, detecting dense facial feature points for profile faces is difficult due to the severe self-occlusion of the face, which in turn restricts the range of poses that these methods can handle. In stead, only the five most stable facial feature points are utilized for pose normalization in this chapter. This greatly facilitates the realization of a fully automatic face recognition system and extends the range of poses that can be processed. Although using sparse facial feature points will result in larger normalization error, we highlight the power of the proposed PBPR-MtFTL framework given its low normalization requirements.

4.3.2 Unoccluded Facial Texture Detection

Pose normalization corrects the deformation of facial texture resulting from pose variations, but it cannot recover the texture lost by occlusion. Rather than trying to synthesize the occluded texture to obtain a complete frontal face [Abiantun et al., 2014], we propose to make full use of the unoccluded texture only. This is inspired by the observation that human beings can easily recognize profile faces without the need

[1]In this step, we roughly estimate the pose of the 2D face image by the method described in [Bruckstein et al., 1999].

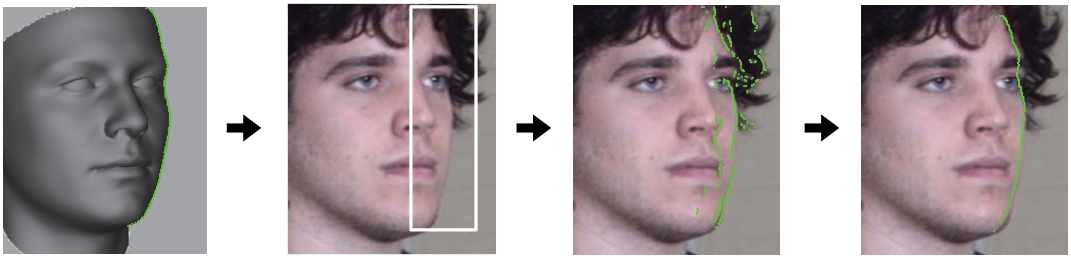
to recover the whole frontal face. As shown in Fig. 4.3, the main boundary between the occluded and unoccluded facial texture is the facial contour. Therefore, facial contour detection is the key to identifying the occluded facial texture.

Although there are off-the-shelf face alignment tools for facial contour detection, they return only sparse facial contour points and may not be reliable enough to severe occlusion, expression, and pose variations. We propose a much simpler but effective method that makes use of the 3D generic shape model. After aligning the 3D model and the 2D face image, it can be projected to the 2D image plane roughly in the pose of the 2D face. As shown in Fig. 4.4(a), the contour of the 3D model can be easily detected. Based on the contour of the 3D model, the facial contour search of the 2D face can be constrained within a certain region, as illustrated in Fig. 4.4(b). The edge points are then detected in this region by the Canny operator [Canny, 1986]. To reduce imposters, only the edge points with horizontal gradient directions are saved, with the prior that facial contour extends in a vertical direction. Lastly, the facial contour is obtained by a point sets registration algorithm called Coherent Point Drift (CPD) [Myronenko and Song, 2010]. Briefly, CPD iteratively aligns the facial contour highlighted in Fig. 4.4(a) to the edge point set shown in Fig. 4.4(c) with affine transformations. The imposter contour points in Fig. 4.4(c) can gradually be detected and ignored. The obtained facial contour is shown in Fig. 4.4(d). More detection examples on unconstrained face images in the LFW dataset [Huang et al., 2007] are shown in Fig. 4.5.

Alongside the projection of facial texture in the first step, the detected facial contour points are first projected to the 3D model and then projected to the rendered frontal face image. Since the head is approximately an ellipsoid, the facial contour points in the frontal view are fitted with an arc. As shown in Fig. 4.3, the arc effectively separates the unoccluded and occluded texture in the rendered frontal image. In the following subsection, face representation is built using only the detected unoccluded facial texture.

4.3.3 Patch-based Face Representation

The area of the unoccluded facial texture in the rendered frontal view varies with pose change, with demonstrable fluctuation in the amount of effective information available for face recognition. In light of this observation, a variable-length face representation



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

Figure 4.4: Illustration of facial contour detection. (a) The 3D generic shape model is projected to the 2D plane and its facial contour is detected; (b) the region containing the facial contour of the 2D face image is estimated; (c) candidate facial contour points; (d) facial contour obtained by point set registration.

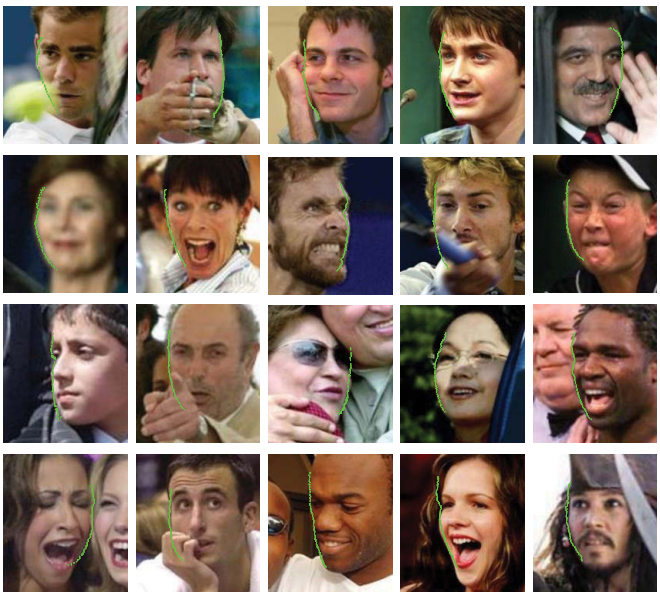


Figure 4.5: Examples of facial contour detection for unconstrained face images in the LFW dataset.

method is proposed.

As illustrated in Fig. 4.3, the normalized face image is first divided into $M \times N$ overlapped patches. The severity of occlusion for each patch is then evaluated based on the detected boundary between the occluded and unoccluded facial texture. If more than 80% of pixels in one patch fall into the unoccluded region, then it is designated as an unoccluded patch; otherwise, the patch is ignored due to the large area of occlusion. Next, each of the unoccluded patches is split into $J \times J$ cells. A state-of-the-art local descriptor called Dual-Cross Patterns (DCP) [Ding et al., 2016] is employed for feature extraction. The concatenated DCP histogram feature from the J^2 cells forms the raw feature of the patch. Following [Ding et al., 2016], elements in the DCP histogram are normalized by square root. Lastly, Principal Component Analysis (PCA) is applied to each patch to project its feature into a subspace with dimension D, by which the noise is suppressed.

The set of patch-level DCP features following PCA processing from all unoccluded patches forms the representation of the face image. Note that this representation method is general in nature, meaning that it applies to faces with arbitrary poses. This is a valuable property, because we do not need to apply different algorithms to frontal and non-frontal faces, unlike some existing approaches [Ho and Chellappa, 2013].

4.4 Multi-task Feature Transformation Learning

The previous section has introduced the PBPR face representation scheme, whereby face recognition can be accomplished by directly matching corresponding patch features of two face images. In this section, we further propose the MtFTL approach for learning transformation dictionaries, which enable the patch features of a frontal face and a non-frontal face to be transformed into a common discriminative space to enhance recognition ability. The learning process is patch-wise, which means a separate transformation dictionary is learnt by MtFTL for each patch. Consequently, we obtain $M \times N$ transformations dictionaries. Details of the MtFTL approach are illustrated below.

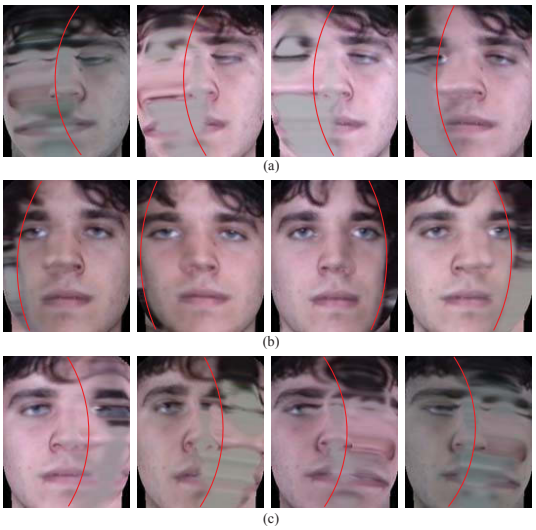


Figure 4.6: Pose normalization for non-frontal images. The boundary between unoccluded and occluded facial texture is detected by the method illustrated in Fig. 4.3. (a) $-90° \leq yaw \leq -45°$; (b) $-30° \leq yaw \leq +30°$; (c) $+45° \leq yaw \leq +90°$. The image quality is degraded with the increase in value of the yaw angles, and the amount of unoccluded facial texture for recognition decreases.

4.4.1 Feature Transformation Learning

Three aspects are considered in the design of feature transformation learning. First, as shown in Fig. 4.6, the normalized images from different poses are of different image quality, therefore there will be differences in the transformations for different poses. Second, a strong correlation exists between the feature transformations for different poses, since they essentially process the data of the same subjects. Third, the amount of training data might be limited in real scenarios, because collecting multi-pose face images tends to be difficult. Ideally, the shared knowledge from different poses should be leveraged for robust transformation learning.

These considerations call for a multi-task strategy for feature transformation learning, in which the learning for each pose type is regarded as a task. Therefore, we propose the MtFTL approach which takes into consideration both the correlation and difference between tasks. Instead of learning a separate transformation matrix for each task [Ma et al., 2014], MtFTL learns a common transformation dictionary for all the tasks. Differences between the tasks are reflected by the selection of

different projection vectors in the transformation dictionary. Hence, MtFTL learns more compact feature transformations than previous approaches [Ma et al., 2014].

Before presenting the formulation of the proposed model, several necessary notations are introduced. Let P be the number of tasks, i.e., the number of pose types that are available in the training set for the current patch. The set $\{(X_t, Y_t) : 1 \leq t \leq P\}$ stores the training data composed of intra-personal and inter-personal patch pairs. $X_t \in \mathbb{R}^{D \times N_{tp}}$ and $Y_t \in \mathbb{R}^{D \times N_{tn}}$, where N_{tp} and N_{tn} are the number of intra-personal and inter-personal patch pairs for the tth task, respectively. The nth column of X_t is denoted as x_t^n and $x_t^n = x_{tn} - x_{0n}$, where $\{x_{tn}, x_{0n}\}$ is one intra-personal patch pair between the pose type t and the frontal pose. Similarly, y_t^n is the nth column of Y_t and $y_t^n = y_{tn} - y_{0n}$, where $\{y_{tn}, y_{0n}\}$ is one inter-personal patch pair between the pose type t and the frontal pose. We learn the transformation dictionary $U \in \mathbb{R}^{D \times D}$ shared by all tasks, and the set of vectors $\alpha_t \in \{0, 1\}^D, 1 \leq t \leq P$. α_t selects projection vectors for task t from the shared dictionary U. Let $A \in \mathbb{R}^{D \times P}$ be the matrix of stacked vectors α_t. In addition, $A_t \in \mathbb{R}^{D \times D}$ is a diagonal matrix expanded by α_t, denoted as $A_t = diag(\alpha_t)$.

The loss function for task t is denoted as $T_{U,t}$. It is based on the principle that the margin between intra-personal patch pairs and inter-personal patch pairs should be as large as possible.

$$T_{U,t}(\alpha_t) = \frac{1}{N_{tp}} \|A_t U^T X_t\|_F^2 - \frac{\lambda}{N_{tn}} \|A_t U^T Y_t\|_F^2, \tag{4.1}$$

where λ is a regularization parameter that weights the intra-personal and inter-personal terms. We then formulate the multi-task learning algorithm as the optimization problem:

$$\begin{aligned} \min_{U,A} \quad & \tfrac{1}{P} \sum_{t=1}^{P} T_{U,t}(\alpha_t) \\ s.t. \quad & U^T U = I. \end{aligned} \tag{4.2}$$

For simplicity, the above problem is relaxed to

$$\min_{U,A} \frac{1}{P} \sum_{t=1}^{P} T_{U,t}(\alpha_t) + \mu \|U^T U - I\|_F^2, \tag{4.3}$$

where μ is another regularization parameter. We set the number of non-zero elements in α_t as d, and aim to optimize α_t to select d most discriminative projection vectors from U for the tth task. The optimal value of μ, d, and λ is estimated through cross validation.

Note that the transformation dictionary U is learnt jointly for all P tasks, which enables knowledge sharing between the tasks. We show in the experiment section that knowledge sharing is especially important when the amount of training data for the tasks is limited.

4.4.2 Iterative Optimization Algorithm

Algorithm 1: Multi-task Feature Transformation Learning

Input: data $\{(X_t, Y_t) : 1 \leq t \leq P\}$, parameters λ, μ, d, ϵ
Output: U, A
Initialize $U^{(0)} := rand(d, d)$, $A^{(0)} := \mathbf{0}$, $s := 0$;
repeat
  $s := s + 1$;
  for $t = 1$ **to** P **do**
    Let $\alpha_t^{(s)} = \arg\min_{\alpha_t} T_{U,t}(\alpha_t)$: Eq. (4.5)
  end
  Let $U^{(s)} = \arg\min_{U} T_A(U)$: Eq. (4.7)
  (via the LBFGS algorithm)
  $\Delta U = U^{(s)} - U^{(s-1)}$
  $\Delta A = A^{(s)} - A^{(s-1)}$
until $\|\Delta U\|_F^2 < \epsilon$ *and* $\|\Delta A\|_F^2 < \epsilon$;

The optimization problem (4.3) of MtFTL is convex in U for fixed A and in A for fixed U. Therefore, we solve this problem by alternately optimizing U and A. The final learning algorithm is summarized in Algorithm 1. The main optimization procedure can be outlined in two steps.

Learning A: With fixed U, the optimization problems for each task decouple. For the tth task, the optimal α_t is obtained:

$$\min_{\alpha_t \in \mathbb{R}^D} T_{U,t}(\alpha_t), \tag{4.4}$$

$$
\begin{aligned}
T_{U,t}\left(\alpha_t\right) &= tr\left(\frac{1}{N_{tp}} A_t U^T X_t X_t^T U A_t - \frac{\lambda}{N_{tn}} A_t U^T Y_t Y_t^T U A_t\right) \\
&= tr\left\{A_t\left(\frac{1}{N_{tp}} U^T X_t X_t^T U - \frac{\lambda}{N_{tn}} U^T Y_t Y_t^T U\right) A_t\right\} \\
&= \alpha_t^T B_t \alpha_t,
\end{aligned} \tag{4.5}
$$

where $tr\left(\cdot\right)$ represents the trace of a matrix; and B_t is a diagonal matrix by directly copying the diagonal elements from the matrix $\frac{1}{N_{tp}} U^T X_t X_t^T U - \frac{\lambda}{N_{tn}} U^T Y_t Y_t^T U$. Since the role of α_t is to select d most discriminative projection vectors for the tth pose type, the elements in α_t that correspond to d smallest diagonal elements in B_t are set as 1 while the other elements in α_t are set as 0.

Learning U: The shared transformation dictionary U couples all the tasks. In this step, U is updated efficiently via the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm. The choice of LBFGS algorithm here is due to both its high efficiency and low memory requirement.

While A is fixed, the optimization problem (4.3) reduces to

$$
\min_{U \in \mathbb{R}^{D \times D}} T_A\left(U\right), \tag{4.6}
$$

$$
\begin{aligned}
T_A\left(U\right) =& \frac{1}{P} \sum_{t=1}^{P} \left(\frac{1}{N_{tp}} \|A_t U^T X_t\|_F^2 - \frac{\lambda}{N_{tn}} \|A_t U^T Y_t\|_F^2\right) \\
&+ \mu \|U^T U - I\|_F^2.
\end{aligned} \tag{4.7}
$$

The derivative of $T_A\left(U\right)$ with respect to U is

$$
\begin{aligned}
\frac{\partial T_A\left(U\right)}{\partial U} =& \frac{2}{P} \sum_{t=1}^{P} \left(\frac{1}{N_{tp}} X_t X_t^T U A_t A_t - \frac{\lambda}{N_{tn}} Y_t Y_t^T U A_t A_t\right) \\
&+ 4\mu\left(U U^T U - U\right).
\end{aligned} \tag{4.8}
$$

With the provided formula for calculating $T_A\left(U\right)$ and $\frac{\partial T_A(U)}{\partial U}$, the optimization problem can be readily solved with the LBFGS algorithm [Schmidt].

Initialization: The transformation matrix U is simply initialized with a random

matrix whose elements are drawn from the standard uniform distribution on the open interval $(0, 1)$. In the experiment section, we show that even the randomly initialized U achieves promising performance.

Stopping criterion: The iterative optimization process stops when the Frobenius norms of both ΔU and ΔA are below $\epsilon = 10^{-3}$, where ΔU and ΔA are the difference matrices between two successive iterations for U and A, respectively.

4.4.3 Theoretical Analysis

In this subsection, we study the robustness and generalization error of the proposed MtFTL algorithm. The detailed proof can be found in Sections 4.8 and 4.9. All through the theoretical analysis, we consider the loss function for face patch feature x_θ at non-frontal pose θ as

$$\ell(A, U, x, \theta) = \|A_\theta U^T (x_\theta - x_0)\|, \tag{4.9}$$

whose maximum value is assumed to be B.

4.4.3.1 Robustness Analysis

If two corresponding face patch features of two images are from the same subject, then their associated losses are close. This property is formalized as "robustness" in [Xu and Mannor, 2012], and the precise definition is given below:

Definition 1. *An algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$ robust, for $K \in \mathbb{N}$ and $\epsilon(\cdot) : \mathcal{Z} \to \mathbb{R}$, if the sample $\mathcal{Z}$ can be partitioned into K disjoint sets, denoted as $\{C_i\}_{i=1}^K$, so that the following holds for all $\mathbf{s} \in \mathcal{Z}$, given the loss function $\ell(\mathcal{A}_s, z)$ of the algorithm $\mathcal{A}_s$ trained on s:*

$$\begin{aligned} &\forall s \in \mathbf{s}, \forall z \in \mathcal{Z}, \forall i = 1, \cdots, K : \\ &if\ s, z \in C_i, then\ |\ell(\mathcal{A}_s, s) - \ell(\mathcal{A}_s, z)| \leq \epsilon(\mathbf{s}). \end{aligned}$$

Given two face patch features s and z of the same subject from different poses θ_s and θ_z, if $\|s - z\| \leq \gamma$ and $|\theta_s - \theta_z| \leq \Delta_\theta$, we suggest that these two face patch features are close. We assume that for one subject, the difference between any of its

non-frontal face patch feature x_θ and its frontal face patch feature x_0 can be bounded by $\|x_\theta - x_0\| \leq \gamma_0$. Since matrix A_θ at pose θ is a sparse diagonal matrix that has d non-zero elements, we have $\|A_\theta\| \leq \sqrt{d}$. Also, we restrict $\|A_{\theta_1} - A_{\theta_2}\| \leq \Omega_{\Delta_\theta}$ for any two matrices A_{θ_1} and A_{θ_2} at different poses. A face recognition algorithm is said to be robust if the corresponding face patch features from images of the same subject have close losses. This robustness can be measured by the following theorem.

Theorem 1. *Example z is in space $\mathcal{Z} \subset \mathbb{R}^D$, which can be partitioned into K disjoint sets and denoted as $\{C_{i=1}\}_{i=1}^K$. Given the algorithm $\mathcal{A}$ $\{A, U : z \to \mathbb{R}^d\}$, we have for any $\boldsymbol{s} \subset \mathcal{Z}$,*

$$|\ell(\mathcal{A}_s, z) - \ell(\mathcal{A}_s, s)| \leq \sqrt{d}\gamma + \Omega_{\Delta_\theta}\gamma_0$$
$$\forall\ i, j = 1, \cdots, K\ : s \in C_i \text{ and } z \in C_j.$$

Hence $\mathcal{A}$ is $(K, \sqrt{d}\gamma + \Omega_{\Delta_\theta}\gamma_0)$-robust.

Robustness is a fundamental property which ensures that a learning algorithm performs well. Since the sparse diagonal matrices A_s and A_z, which select projection vectors from the transformation dictionary U, are learned from face patches of different poses, they cannot be identical. According to Theorem 1, it is instructive to suggest that the robustness of the algorithm will be improved for the face patches at close poses if their feature transformations have more shared elements, that is, encouraging Ω_{Δ_θ} to be small.

4.4.3.2 Generalization Analysis

Based on the robustness analysis, we show a PAC generalization bound for the algorithm, i.e., the difference between the expected error $\mathcal{L}(\mathcal{A}_s)$ and the empirical error $\mathcal{L}_{emp}(\mathcal{A}_s)$. We begin by presenting a concentration inequality [Van Der Vaart and Wellner, 1996] that helps to derive the bound.

Proposition 1. *Let $(|N_1|, \cdots, |N_K|)$ be an IID multinomial random variable with parameters n and $(\beta(C_1), \cdots, \beta(C_K))$. By the Breteganolle-Huber-Carol inequality we have $Pr\{\sum_{i=1}^K |\frac{N_i}{n} - \beta(C_i)| \geq \zeta\} \leq 2^K \exp(\frac{-n\zeta^2}{2})$, hence with probability at least*

$1-\delta$,

$$\sum_{i=1}^{K}|\frac{N_i}{n}-\beta(C_i)|\leq\sqrt{\frac{2K\ln 2+2\ln(1/\delta)}{n}}.$$

The generalization error bound is presented in the following theorem.

Theorem 2. *If the algorithm $\mathcal{A}$ is $(K,\epsilon(\cdot))$-robust and the training sample $\boldsymbol{s}$ is composed of n examples $\{s_i\}_{i=1}^{n}$, which are generated from β, then for any $\delta>0$, with the probability at least $1-\delta$ we have,*

$$|\mathcal{L}(\mathcal{A}_s)-\mathcal{L}_{emp}(\mathcal{A}_s)|\leq\epsilon(s)+B\sqrt{\frac{2K\ln 2+2\ln(1/\delta)}{n}}.$$

By combining the results of Theorem 1 and Theorem 2, we can easily illustrate the generalization error of the proposed algorithm. Exploiting the shared information of face patches from different poses can strengthen the robustness of the algorithm and then improve the generalization error.

4.5 Face Matching with PBPR-MtFTL

In this section, the face matching problem is addressed based on the proposed PBPR-MtFTL framework. It is assumed that $\{(U^i,A^i):1\leq i\leq MN\}$, i.e., the set of patch-wise transformation dictionaries and selection matrices, has been learnt by MtFTL.

Suppose we are matching a probe face image x_t of pose type t to a frontal gallery face image x_0. It is also assumed that there are K unoccluded patches for x_t. Without loss of generality, we denote the sets of features for the K patches as $\{x_{t1},x_{t2},\cdots,x_{tK}\}$ and $\{x_{01},x_{02},\cdots,x_{0K}\}$ for x_t and x_0, respectively. First, the features of each patch pair $\{(x_{tk},x_{0k}):1\leq k\leq K\}$ are projected into the discriminative space using the learnt U^k and A^k.

$$\begin{aligned}\hat{x}_{tk}&=A_t^k(U^k)^Tx_{tk}\\\hat{x}_{0k}&=A_t^k(U^k)^Tx_{0k},\end{aligned}\tag{4.10}$$

where A_t^k is the diagonal matrix expanded by the tth column of A^k. Then, the cosine metric is utilized to calculate the similarity of each patch pair and the similarity scores

of all K patch pairs are fused by the sum rule.

$$s(x_t, x_0) = \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{x}_{tk}^T \hat{x}_{0k}}{\|\hat{x}_{tk}\| \|\hat{x}_{0k}\|}, \tag{4.11}$$

where s is the similarity score between the probe image x_t and the gallery image x_0. Lastly, the nearest neighbor (NN) classifier is adopted for face identification.

Although the adopted matching scheme is simple compared to existing methods [Li et al., 2009, 2012a], it is still expected that the proposed PBPR-MtFTL framework will achieve stronger performance, since the recognition ability of PBPR-MtFTL has been enhanced by exploiting the correlation between poses.

4.6 Experimental Evaluation

In this section, extensive experiments are conduct to present the effectiveness of PBPR-MtFTL. We mainly conduct identification experiments on the three most popular databases for the pose problem, i.e., CMU-PIE [Sim et al., 2003], FERET [Phillips et al., 2000], and Multi-PIE [Gross et al., 2010]. These experiments are to recognize a subject across pose variations with a single enrolled frontal face image. At the end of this section, we slightly modify the proposed framework to deal with the unconstrained face verification problem, and conduct experiments on the challenging LFW dataset [Huang et al., 2007].

The CMU-PIE [Sim et al., 2003] and FERET [Phillips et al., 2000] datasets incorporate multi-pose images of 68 and 200 subjects, respectively. For the two databases, we adopt the same protocols as previous works [Arashloo and Kittler, 2011, Li et al., 2012c] that exclude both illumination and expression variations. The Multi-PIE [Gross et al., 2010] database contains images of 337 subjects, each of which is captured in up to four recording sessions. Images in each session cover 15 view points and 20 illumination conditions. As there is no unified protocol for the pose problem on Multi-PIE, we adopt the three most popular protocols in the literature [Asthana et al., 2011, Li et al., 2012a, Zhu et al., 2014b].

Eight sets of experiments are conducted. First, the performance of PBPR-MtFTL is briefly compared with previous works for PIFR on CMU-PIE and FERET. Next, the

MtFTL approach is compared with its single-task baselines on Multi-PIE to justify the significance of MTL for the pose problem. Then, considering that the pose problem is often combined with other factors, we evaluate the performance of PBPR-MtFTL in three different settings, i.e., combined variations of pose and illumination, combined variations of pose and recording session, and combined variations of pose, illumination, and recording session. We also test the sensitivity of PBPR-MtFTL to the value of model parameters and face alignment errors. Lastly, we slightly modify the proposed approach to deal with the unconstrained face verification problem and present experimental results on the LFW database.

All images in this chapter are normalized as follows. The mean shape of the Basel Face Model (BFM) [Paysan et al., 2009] is adopted as the 3D generic shape model. The five facial feature points are manually labeled in the first six experiments and automatically detected in the last two experiments[1]. After the pose normalization step described in Section 4.3, the face images are cropped and resized to 156×130 pixels, as shown in Fig. 4.6. The patch size $M \times N$ is set at 26×24 pixels, with 50% overlap between nearby patches. The number of cells $J \times J$ within each patch is set at 2×2. For the first seven experiments, images are further photometrically normalized using a simple operator [Tan and Triggs, 2010], with the two parameters σ_1 and σ_2 set at 1.4 and 2.0. The two parameters R_{in} and R_{ex} for DCP are set at $[3, 7]$. For the last experiment, we omit the photometric normalization step since it slightly degrades the performance of PBPR-MtFTL on the View 1 data of LFW. Three-scale DCP features are extracted and concatenated, with the parameters set at $[2, 4]$, $[4, 8]$, and $[6, 12]$, respectively.

For each identification experiment, the subjects in the training data are randomly divided into two subsets, one for model training and the other for validation. The two subsets are of equal size. The optimal values of model parameters μ, d, and λ are estimated on the validation subset and applied to the test data. For simplicity, the value of μ across the models of all patches keeps consistent and this applies to d and λ. The random division of the training data is repeated five times, and the mean rank-1 identification rates on the test data are reported. The estimated model parameters for different databases are tabulated in Table 4.1, where the superscripts " $*^1$ ", " $*^2$ ",

[1]Coordinates of manually labeled facial feature points for Multi-PIE are provided by. Zhu et al. [2013]; we use an off-the-shelf tool [Sun et al., 2013a] for automatic facial feature point detection.

Table 4.1: Model Parameters Estimated on the Validation Subsets for Different Databases

Database	D	μ	d	λ
CMU-PIE	300	0.1	200	0.5
FERET	200	0.1	50	0.5
Multi-PIE[1]	300	0.1	200	0.5
Multi-PIE[2]	300	0.1	200	0.5
Multi-PIE[3]	300	0.1	200	0.8
LFW	600	0.05	200	0.8

and " $*^3$ " stand for one of the three protocols adopted for the Multi-PIE database, respectively.

4.6.1 Comparison on CMU-PIE and FERET

All 68 CMU-PIE subjects with neutral expression and normal illumination at 11 different poses are employed. Note that pose types C31 and C25 are with hybrid yaw and pitch variations. The 68 frontal images are utilized as gallery images and all the rest are used as probes. Following previous works [Li et al., 2012c, 2014b], we train the MtFTL model with randomly selected 50 subjects in Multi-PIE database since there are only 68 subjects in CMU-PIE. For FERET, all 200 subjects at 9 different poses are incorporated. Images of the first 100 subjects consist the training data and the rest 100 subjects are used for testing.

As shown in Table 4.2 and 4.3, the proposed PBPR-MtFTL approach outperforms the other methods. But the advantage of PBPR-MtFTL is not well exhibited, because the performance of existing methods has nearly reached the saturation point on the two databases. Therefore, we focus on the larger and more challenging Multi-PIE database in the following experiments.

4.6.2 Comparison with Single-task Baselines

In this experiment, we aim to justify the importance of MTL for the pose problem. The proposed MtFTL algorithm is compared with three single-task baselines: (a) the

Table 4.2: Performance Comparison with State-of-the-art PIFR Methods on CMU-PIE

	C31	C34	C14	C11	C29	C05	C37	C02	C22	C25
yaw angle	$-47°$	$-66°$	$-47°$	$-32°$	$-17°$	$+16°$	$+31°$	$+44°$	$+62°$	$+44°$
pitch angle	$+11°$	$0°$	$0°$	$0°$	$0°$	$0°$	$0°$	$0°$	$0°$	$+11°$
SA-EGFC [Li et al., 2012c]	-	91.00	100.0	100.0	100.0	100.0	100.0	100.0	96.00	-
Multi-scale MRFs [Arashloo et al., 2011]	88.00	83.00	91.00	89.00	91.00	98.00	100.0	95.00	79.00	95.00
MRFs+DAISY [Arashloo and Kittler, 2014]	100.0	95.00	100.0	100.0	100.0	100.0	100.0	100.0	92.00	100.0
PBPR-MtFTL	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.53	100.0

Table 4.3: Performance Comparison with State-of-the-art PIFR Methods on FERET

Methods	Gallery Size	bi $-65°$	bh $-45°$	bg $-30°$	bf $-15°$	be $+15°$	bd $+30°$	bc $+45°$	bb $+65°$	Mean
VAAM [Asthana et al., 2011]	200	-	90.50	98.00	98.50	97.50	97.00	91.90	-	-
MRFs+MLBP [Arashloo, 2013]	200	92.00	98.50	99.50	100.0	99.50	99.00	99.50	91.00	97.38
RR+Gabor [Li et al., 2012a]	100	78.00	91.00	96.00	96.00	98.00	99.00	96.00	87.00	92.63
SA-EGFC [Li et al., 2012c]	100	98.00	100.0	100.0	100.0	100.0	100.0	100.0	98.00	99.50
PBPR-MtFTL	100	100.0	99.00	100.0	100.0	100.0	100.0	100.0	98.00	99.63

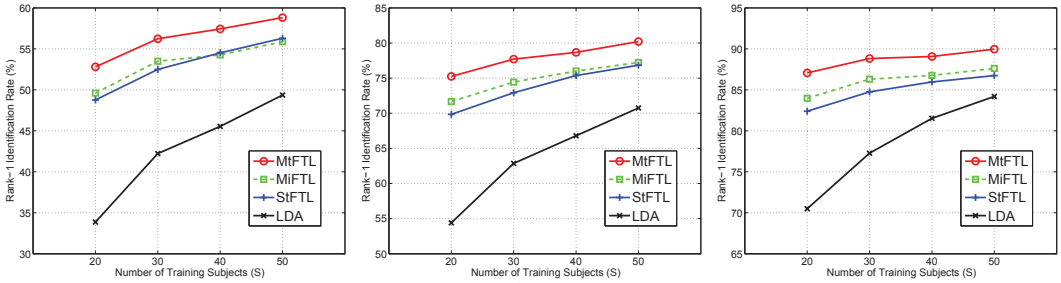


Figure 4.7: Performance comparison of MtFTL and the three single-task baselines on the Multi-PIE database with varying numbers of training subjects. (a) yaw = $\pm 90°$; (b) yaw = $\pm 75°$; (c) yaw = $\pm 60°$.

Linear Discriminative Analysis (LDA) approach, which learns a single LDA model for all poses; (b) the single-task feature transformation learning (StFTL) approach, which learns a single feature transformation for all poses. StFTL is equal to the Discriminative Locality Alignment (DLA) model [Zhang et al., 2009]; (c) the multiple independent feature transformation learning (MiFTL) approach, which independently learns a DLA model for each pose. Unlike LDA and StFTL, MtFTL and MiFTL learn pose specific feature transformations. The main difference between MtFTL and MiFTL is that MiFTL learns the transformation for each pose independently, while MtFTL learns compact transformations simultaneously and benefits from the correlation of different poses.

The protocol defined in [Li et al., 2012a] is employed. This protocol covers 249 subjects in Session 1, in which images with neutral expression under 20 illumination conditions are involved. The first 100 subjects (Subject ID 001 to 100) are used for training and the remaining 149 subjects (Subject ID 101 to 250) are used for testing. The gallery set is composed of 149 frontal images (Pose ID 051) with the illumination ID 07. The probe sets cover 20 illumination conditions of the same subjects. In Fig. 4.7, we present the performance of the four algorithms with varied size of training data on the three most challenging poses. The number of subjects (S) utilized for model learning is gradually increased from 20 to the maximum number 50.

It is shown in Fig. 4.7 that MtFTL consistently outperforms the baselines under all settings. Specifically, MtFTL significantly outperforms StFTL and LDA, which means that learning pose specific feature transformations is necessary. MtFTL outperforms

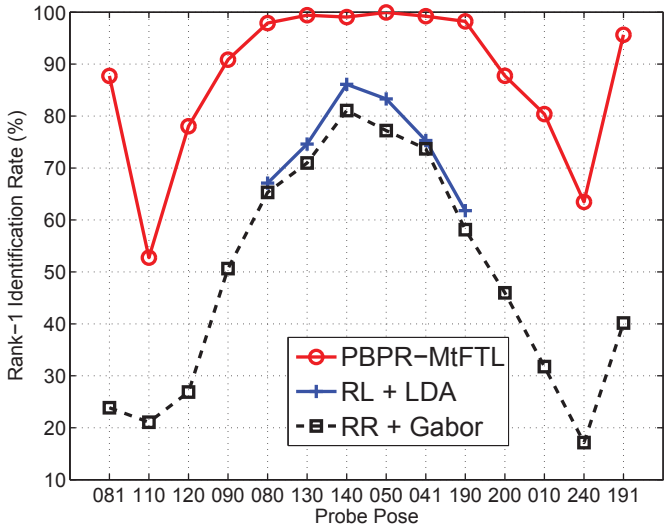


Figure 4.8: Performance comparison on combined variations of pose and illumination. The probe sets 081 and 191 are with hybrid yaw and pitch variations. The other probe sets contain only yaw variations from $-90°$ to $+90°$.

MiFTL while learning much more compact transformations. This proves that MTL is helpful for enhancing the ability to recognize non-frontal faces. The advantage of MtFTL is more evident when the amount of training data is limited, which indicates that knowledge sharing among related tasks is important for better generalization ability.

4.6.3 Recognition across Pose and Illumination

In this subsection, the performance of the PBPR-MtFTL framework is compared with existing algorithms under the setting of the combined variations of pose and illumination. The adopted protocol is the same as the previous experiment [Li et al., 2012a]. The experimental results are shown in Table 4.4 and Fig. 4.8. In general, face recognition across combined variations of pose and illumination is a difficult problem. However, it is clear that the proposed method outperforms existing approaches [Li et al., 2012a, Zhu et al., 2013] with a large margin, even though only half training data is employed to train the MtFTL model.

It is worth noting that the algorithm proposed in [Li et al., 2012a] also employs photometric normalization, and that all three approaches employ manually labeled

facial feature points. Notably, we employ exactly the same facial feature point coordinates as the method followed in [Zhu et al., 2013].

4.6.4 Recognition across Pose and Recording Session

This experiment is to test the performance of algorithms under the combined variations of pose and recording session. The protocol described in [Asthana et al., 2011] is followed. This protocol covers all 337 subjects across the four recording sessions. Only images with neutral expression and frontal illumination are employed. Images of the first 200 subjects (Subject ID 001 to 200) are used for training, and images of the remaining 137 subjects (Subject ID 201 to 346) are employed for testing. The frontal images from the earliest recording sessions for the testing subjects are collected as the gallery set (137 images in total). The non-frontal images of the testing subjects construct fourteen probe sets. The comparisons between our approach and the state-of-the-art methods are presented in Table 4.5 and Fig. 4.9. We observe that:

1. In general, the performance of all the algorithms is good when the pose value of the probe images is small. While high performance is achieved by all methods on the probe sets 130, 140, 050, and 041, our method achieves perfect identification rates on all four probe sets.

2. There is a substantial drop in performance for existing methods on the probe sets 080 and 190, where the yaw angles are $\pm 45^{\circ}$. PBPR-MtFTL performs significantly better than the other methods on both probe sets, indicating that it is more robust to large pose variations.

3. While most existing methods can only handle yaw angle variations within $[-45^{\circ}, +45^{\circ}]$, the proposed method can tackle the full range of yaw angle variation. Fig. 4.9 shows that high performance is achieved even when the yaw angle approaches $\pm 75^{\circ}$.

Table 4.4: Rank-1 Identification Rates on Combined Variations of Pose and Illumination on Multi-PIE

Methods	090 $-60°$	080 $-45°$	130 $-30°$	140 $-15°$	050 $+15°$	041 $+30°$	190 $+45°$	200 $+60°$	mean
RR+Gabor [Li et al., 2012a]	50.64	65.30	70.97	81.07	77.21	73.69	58.12	45.97	65.37
RL+LDA [Zhu et al., 2013]	-	67.10	74.60	86.10	83.30	75.30	61.80	-	-
PBPR-MtFTL	**90.86**	**97.91**	**99.41**	**99.05**	**99.94**	**99.23**	**98.21**	**87.75**	**96.55**

Table 4.5: Rank-1 Identification Rates on Combined Variations of Pose and Recording Session on Multi-PIE

Methods	Alignment	080 $-45°$	130 $-30°$	140 $-15°$	050 $+15°$	041 $+30°$	190 $+45°$	Mean
VAAM [Asthana et al., 2011]	Auto	74.10	91.00	95.70	95.70	89.50	74.80	86.80
SA-EGFC [Li et al., 2012c]	Manual	93.00	98.70	99.70	99.70	98.30	93.60	97.17
MRFs [Ho and Chellappa, 2013]	N/A	86.30	89.70	91.70	91.00	89.00	85.70	88.90
RL+LDA [Zhu et al., 2013]	Manual	95.60	98.50	**100.0**	99.30	98.50	97.80	98.28
SPAE [Kan et al., 2014]	Auto	84.90	92.60	96.30	95.70	94.30	84.40	91.37
MDF-PM [Li et al., 2014b]	Manual	90.00	94.30	95.30	94.70	93.70	87.70	92.62
MVP+LDA [Zhu et al., 2014b]	Manual	93.40	**100.0**	**100.0**	**100.0**	99.30	95.60	98.05
PBPR-MtFTL	Manual	**98.67**	**100.0**	**100.0**	**100.0**	**100.0**	**98.33**	**99.50**

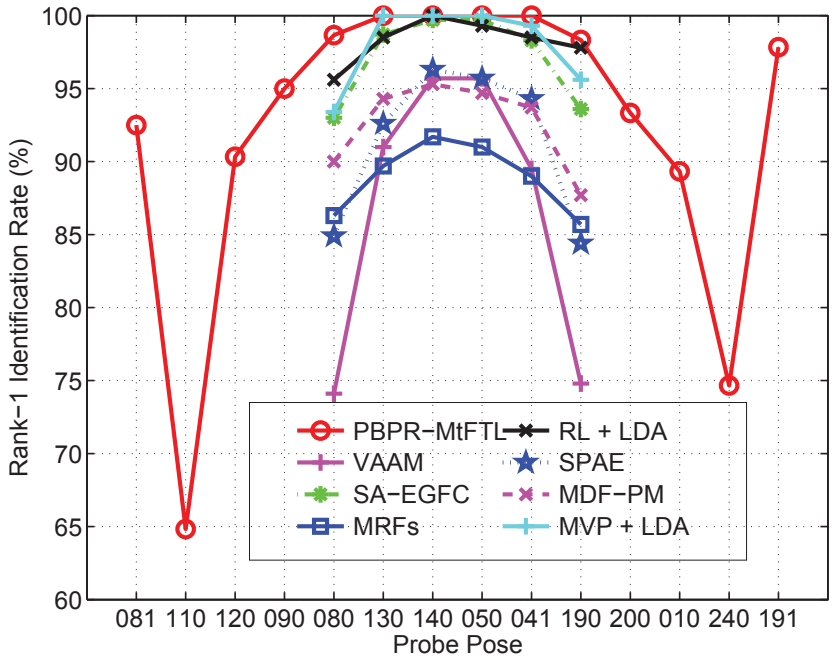


Figure 4.9: Performance comparison of different methods on combined variations of pose and recording session.

4.6.5 Recognition across Pose, Illumination, and Recording Session

To examine the robustness of the proposed algorithm under more challenging conditions, a new protocol specified in [Zhu et al., 2014b] is employed. This protocol extends the original protocol designed in [Asthana et al., 2011] by incorporating all 20 illumination types, while the other settings remain the same. Therefore, the gallery set is exactly the same as [Asthana et al., 2011], while the number of probe images is 20 times more than that in [Asthana et al., 2011]. The performance of the proposed method, compared with the state-of-the-art approaches, is presented in Table 4.6 and Fig. 4.10. All methods in Table 4.6 employ manually labeled facial feature points. We make the following observations:

1. PBPR-MtFTL significantly outperforms the other three approaches across all probe sets. This result is consistent with those observed in the previous two experiments.

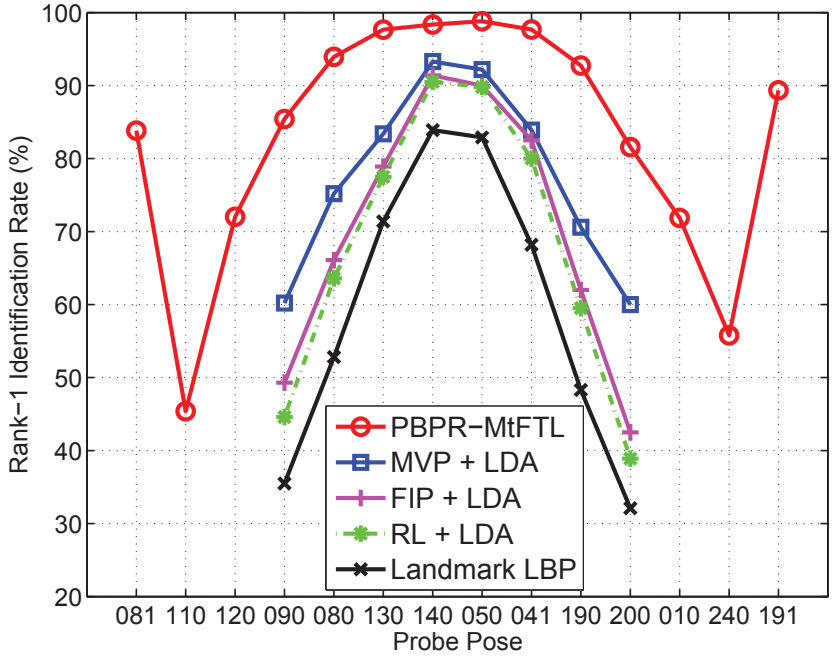


Figure 4.10: Performance comparison of different methods on combined variations of pose, illumination, and recording session.

2. Among the four approaches, PBPR-MtFTL is the only one that can handle full range of pose variations, and its performance degrades gracefully across wide pose variations including $\pm 60°$.

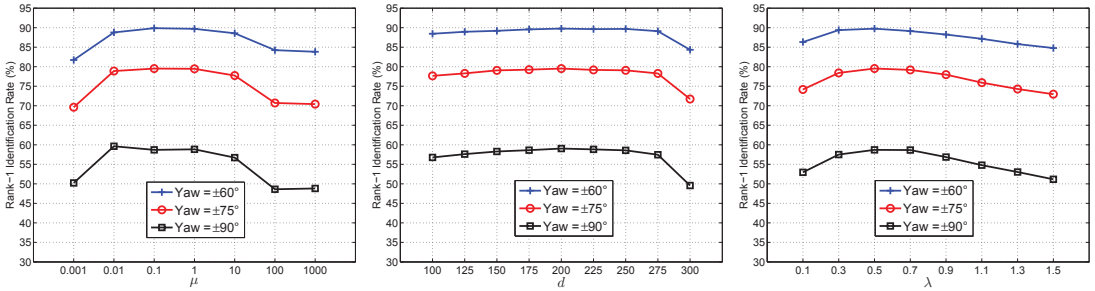
4.6.6 Parameter Evaluation for MtFTL

In the above experiments, the optimal value of model parameters μ, d, and λ is estimated on the validation subsets. In this experiment, the impact of their value on the performance of MtFTL is investigated. The same protocol [Li et al., 2012a] as the second experiment is followed. Also, the rank-1 identification rates on the three most challenging poses are reported.

The performance of the MtFTL approach for different value of μ, d, and λ is shown in Fig. 4.11. Their optimal value is around 0.1, 200, and 0.5, respectively. The experimental results also indicate that the performance of MtFTL is robust to the fluctuation of parameter value.

Table 4.6: Rank-1 Identification Rates on Combined Variations of Pose, Illumination, and Recording Session on Multi-PIE

Methods	090 $-60°$	080 $-45°$	130 $-30°$	140 $-15°$	050 $+15°$	041 $+30°$	190 $+45°$	200 $+60°$	Mean
Landmark LBP+LDA [Chen et al., 2013]	35.50	52.80	71.40	83.90	82.90	68.20	48.30	32.10	59.39
FIP+LDA [Zhu et al., 2013, 2014b]	49.30	66.10	78.90	91.40	90.00	82.50	62.00	42.50	70.34
RL+LDA [Zhu et al., 2013, 2014b]	44.60	63.60	77.50	90.50	89.80	80.00	59.50	38.90	68.05
MVP+LDA [Zhu et al., 2014b]	60.20	75.20	83.40	93.30	92.20	83.90	70.60	60.00	77.35
PBPR-MtFTL	**85.41**	**93.93**	**97.66**	**98.36**	**98.81**	**97.68**	**92.74**	**81.58**	**93.27**



(a) (b) (c)

Figure 4.11: Influence of the parameters μ, d, and λ to the performance of MtFTL. (a) evaluation against the value of μ while d and λ are set at 200 and 0.5, respectively; (b) evaluation against the value of d while μ and λ are set at 0.1 and 0.5, respectively; (c) evaluation against the value of λ while μ and d are set at 0.1 and 200, respectively.

4.6.7 Performance in the Fully-Automatic Mode

The performance of the presented PBPR-MtFTL framework is related to the accuracy of the facial feature detection and pose estimation algorithms. The previous experiments are semi-automatic (SA), i.e., the facial feature points are labeled manually and it is assumed that the probe image poses are known. In this experiment, the PBPR-MtFTL framework is run in the fully-automatic (FA) mode. We leave the manually labeled facial feature points for the gallery and training images intact. This is reasonable since the labeling work could be conducted offline. For all probe images, the five facial feature points are automatically detected. Since existing face alignment tools cannot reliably detect facial feature points for profile or half-profile faces, we limit the yaw range of the probe images to within $\pm 45^\circ$ in this experiment. For pose estimation, we compare the unoccluded region of each probe image with those of a set of training images whose poses are known. The pose of the probe image is assigned to that of the training image whose unoccluded region is the most similar.

The same protocol [Li et al., 2012a] as used in the second experiment is adopted. The performance of PBPR-MtFTL in the SA and FA modes is compared in Fig. 4.12. There is a minor drop in performance under the FA mode. In fact, the performance drop is mainly caused by the failure of face detection, whose failure rates on the six probe sets are 3.66%, 1.95%, 1.21%, 1.51%, 1.98%, and 3.72%, respectively. Besides, when considered along with the results shown in Fig. 4.8, the performance of PBPR-MtFTL

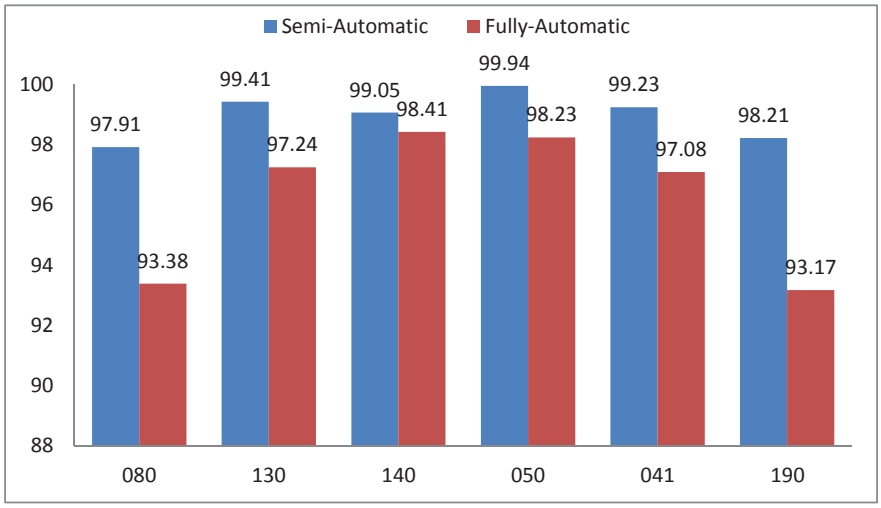


Figure 4.12: Performance comparison of the proposed PBPR-MtFTL framework in the SA and FA modes. In the FA mode, both facial feature point detection and pose estimation are completely automatic. Note that the identification error in the FA mode incorporates the failure in face detection.

in the FA mode is still considerably better than the state-of-the-art methods.

4.6.8 Extension to Unconstrained Face Verification

In this experiment, we slightly modify the proposed approach to tackle the unconstrained face verification problem, and present experimental results on the LFW database [Huang et al., 2007].

In Section 4.4, we assume that the tth task of MtFTL is to learn the feature transformation between the tth non-frontal pose type and the frontal pose. As shown in Fig. 4.13, image pairs defined in LFW may contain no frontal pose image. Therefore, we add tasks in the model that learn the feature transformation between every possible pair of poses. Another characteristic of LFW is that it has very few profile face images. To make sure that each pose type incorporates sufficient training data, we quantize the pose space into three types, i.e., left profile (LP, $yaw < -10°$), frontal pose (FP, $-10° \leq yaw \leq +10°$), and right profile (RP, $yaw > +10°$). Therefore, there are six tasks in total, i.e., LP-LP, LP-FP, LP-RP, FP-FP, FP-RP, and RP-RP. Besides, the weight of each task in Eq. 4.3 is modified to be proportional to the number of training samples in each task, since the training samples are far from balanced among the tasks.

Figure 4.13: Many image pairs defined in LFW contain no frontal faces. The first line shows the first images in the image pairs, while the second line shows the second images in the image pairs.

For each pair of faces in LFW, both images are normalized as described in Section 4.3. Occlusion detection is conducted for non-frontal face images. Features are extracted only from the patches that are un-occluded in both images. The features are transformed with the learnt MtFTL model, as described in Section 4.4. Similarity scores between all un-occluded patch pairs are averaged as the similarity score of the face image pair. Since MtFTL explicitly employs the image labels, the proposed method falls in the paradigm of "Unrestricted, Label-Free Outside Data" [Huang et al., 2007]. We conduct performance comparison with the state-of-the-arts in Table 4.7. To promote performance, most existing methods designed for the LFW challenge fuse multiple face representations, e.g., employing several descriptors [Li et al., 2013a] or mirroring the face image [Simonyan et al., 2013]. In comparison, this work is not targeted at the LFW challenge and we employ only a single face representation. For fair comparison, we report the best performance of all approaches in Table 4.7 achieved with a single face representation[1].

The first five approaches in Table 4.7 adopt metric learning based classifiers, and PBPR-MtFTL achieves significantly better performance than the other approaches. Recently, generative model based classifiers have been introduced to the LFW challenge. We then replace the MtFTL model with the Probabilistic Linear Discriminative Model (PLDA) [Prince and Elder, 2007]. The dimension of the PLDA subspace is set

[1] The performance of [Simonyan et al., 2013] is obtained using the code and data released by the authors, while the performance of the other approaches is directly cited from the original papers.

Table 4.7: Performance comparison on LFW with state-of-the-art methods based on single face representation

Methods	Accuracy ± Error(%)
MOSS + ITML [Taigman et al., 2009]	85.17 ± 0.61
Sub-SML [Cao et al., 2013]	87.15 ± 0.56
OCLBP + LDA [Barkan et al., 2013a]	88.75 ± 0.59
FVF + Mah. Metric [Simonyan et al., 2013]	88.85 ± 2.35
PBPR + MtFTL	**91.78 ± 0.58**
LBP + PLDA [Li et al., 2012b]	87.33 ± 0.55
High-dim LBP + Joint Bayesian [Chen et al., 2013]	**93.18 ± 1.07**
PBPR + PLDA	**92.95 ± 0.37**

at 100. With generative model based classifiers, the high-dim LBP approach [Chen et al., 2013] achieves a slightly higher accuracy than our approach. However, this approach relies on dense facial feature detection. We emphasize here that only the 5 most stable facial feature points are required by our method. This makes our algorithm easier to use in practical applications.

4.7 Conclusion

Face recognition across pose is a challenging task because of the significant appearance change caused by pose variations. We handle this problem from two aspects. First, we propose the PBPR face representation scheme that makes use of the unoccluded face textures only. PBPR can be applied to face images in arbitrary pose, which is a great advantage over existing methods. Second, we present the MtFTL model for learning compact feature transformations by utilizing the correlation between poses. Clear advantage is shown compared to single-task based methods. To the best of our knowledge, this is the first time that MTL has been formally applied to the PIFR problem. As the proposed PBPR-MtFTL framework effectively utilizes all the unoccluded face texture and the correlation between different poses, very encouraging results for face identification in all three popular multi-pose databases are achieved. We also slightly modify the proposed approach to tackle the unconstrained face verification

problem, and achieve top level performance on the challenging LFW database.

4.8 Proof of Theorem 1

Proof. We can partition $\mathcal{Z}$ into K disjoint sets, so that if two face patch features s and z are close, then

$$\|s - z\| \leq \gamma \quad \texttt{and} \quad |\theta_s - \theta_z| \leq \Delta_\theta. \tag{4.12}$$

By arranging the loss functions so that the first loss is always larger than the second one, we therefore have

$$\begin{aligned}
&|\ell(A_s, U, s) - \ell(A_z, U, z)| \\
&= \|A_s U^T(s - x_0)\| - \|A_z U^T(z - x_0)\| \\
&= \|A_s U^T(s + z - z - x_0)\| - \|A_z U^T(z - x_0)\| \\
&\leq \|A_s U^T(s - z)\| + \|A_s U^T(z - x_0)\| - \|A_z U^T(z - x_0)\| \\
&\leq \|A_s U^T(s - z)\| + \|(A_s - A_z + A_z)U^T(z - x_0)\| \\
&\quad - \|A_z U^T(z - x_0)\| \\
&\leq \|A_s U^T(s - z)\| + \|(A_s - A_z)U^T(z - x_0)\| \\
&\leq \|A_s\|\|s - z\| + \|A_s - A_z\|\|z - x_0\| \\
&\leq \sqrt{d}\gamma + \Omega_{\Delta_\theta}\gamma_0,
\end{aligned}$$

which completes the proof. □

4.9 Proof of Theorem 2

Proof. Let N_i be the set of index of points of $\mathbf{s}$ that fall into C_i. $(|N_1|, \cdots, |N_K|)$ is an IID random variable with parameters n and $(\beta(C_1), \cdots, \beta(C_K))$. We have

$$\begin{aligned}
&|\mathcal{L}(\mathcal{A}_s) - \mathcal{L}_{emp}(\mathcal{A}_s)| \\
&= \left| \sum_{i=1}^{K} E_{z\sim\beta}(\ell(\mathcal{A}_s, z)|z \in C_i)\beta(C_i) - \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}_s, s_i) \right| \\
&\leq \left| \sum_{i=1}^{K} E_{z\sim\beta}(\ell(\mathcal{A}_s, z)|z \in C_i)\frac{N_j}{n} - \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}_s, s_i) \right|
\end{aligned}$$

$$
\begin{aligned}
&+ \left| \sum_{i=1}^{K} E_{z\sim\beta}(\ell(\mathcal{A}_s, z)|z \in C_i)\beta(C_i) - \sum_{i=1}^{K} E_{z\sim\beta}(\ell(\mathcal{A}_s, z)|z \in C_i)\frac{N_j}{n} \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^{K} \sum_{j\in N_i} \max_{z\in C_i} |\ell(\mathcal{A}_s, s_j) - \ell(\mathcal{A}_s, z)| \right| \\
&\quad + \left| \max_{z\in\tilde{\mathcal{Z}}} |\ell(\mathcal{A}_s, z)| \sum_{i=1}^{K} |\frac{|N_i|}{n} - \beta(C_i)| \right| \\
&\leq \epsilon(s) + B \sum_{i=1}^{K} |\frac{N_i}{n} - \beta(C_i)| \\
&\leq \epsilon(s) + B\sqrt{\frac{2K\ln 2 + 2\ln(1/\delta)}{n}}.
\end{aligned}
$$

The first inequality is due to the triangle inequality, and the second inequality is because of $\sum_{i=1}^{K} \beta(C_i) = 1$ and $\sum_{i=1}^{K} \frac{N_j}{n} = 1$. Finally, the last inequality is the application of Proposition 1. □

Chapter 5

Trunk-Branch Ensemble Convolutional Neural Networks for Video-based Face Recognition

Human faces in surveillance videos often suffer from severe image blur, dramatic pose variations, and occlusion. In this chapter, we propose a comprehensive framework based on Convolutional Neural Networks (CNN) to overcome challenges in video-based face recognition (VFR). First, to learn blur-robust face representations, we artificially blur training data composed of clear still images to account for a shortfall in real-world video training data. Using training data composed of both still images and artificially blurred data, CNN is encouraged to learn blur-insensitive features automatically. Second, to enhance robustness of CNN features to pose variations and occlusion, we propose a Trunk-Branch Ensemble CNN model (TBE-CNN), which extracts complementary information from holistic face images and patches cropped around facial components. TBE-CNN is an end-to-end model that extracts features efficiently by sharing the low- and middle-level convolutional layers between the trunk and branch networks. Third, to further promote the discriminative power of the representations learnt by TBE-CNN, we propose an improved triplet loss function. Systematic experiments justify the effectiveness of the proposed techniques. Most impressively, TBE-CNN achieves state-of-the-art performance on three popular video face databases: PaSC, COX Face, and YouTube Faces. With the proposed techniques,

we obtain the first place in the BTAS 2016 Video Person Recognition Evaluation.

5.1 Introduction

With the widespread use of video cameras for surveillance and in mobile devices, an enormous quantity of video is constantly being captured. Compared to still face images, videos usually contain more information, e.g., temporal and multi-view information. The ubiquity of videos offers society far-reaching benefits in terms of security and law enforcement. It is highly desirable to build surveillance systems coupled with face recognition techniques to automatically identify subjects of interest. Unfortunately, the majority of existing face recognition literature focuses on matching in still images, and video-based face recognition (VFR) research is still in its infancy [Best-Rowden et al., 2013, Huang et al., 2015b]. In this chapter, we handle the still-to-video (S2V), video-to-still (V2S), and video-to-video (V2V) matching problems, which are used in the most common VFR applications.

Compared to still image-based face recognition (SIFR), VFR is significantly more challenging. Images in standard SIFR datasets are usually captured under good conditions or even framed by professional photographers, e.g., in the Labeled Faces in the Wild (LFW) database [Huang et al., 2007]. In comparison, the image quality of video frames tends to be significantly lower and faces exhibit much richer variations (Fig. 5.1) because video acquisition is much less constrained. In particular, subjects in videos are usually mobile, resulting in serious motion blur, out-of-focus blur, and a large range of pose variations. Furthermore, surveillance and mobile cameras are often low-cost (and therefore low-quality) devices, which further exacerbates problems with video frame clarity [Biswas et al., 2013].

Recent advances in face recognition have tended to ignore the peculiarities of videos when extending techniques from SIFR to VFR [Parkhi et al., Schroff et al., 2015, Sun et al., 2015]. On the one hand, a major difficulty in VFR, such as severe image blur, is largely unsolved [Beveridge et al., 2015]. One important reason is that large amounts of real-world video training data are still lacking, and existing still image databases are usually blur-free. On the other hand, although pose variations and occlusion are partially solved in SIFR by ensemble modelling [Liu et al., 2015, Sun et al., 2015], the strategy may not be directly extended to VFR. A common practice

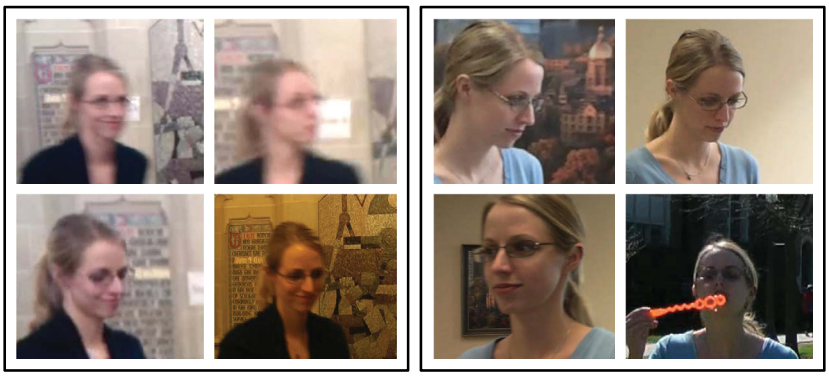


Figure 5.1: Video frames captured by surveillance or mobile devices suffer from severe image blur, dramatic pose variations, and occlusion. (a) Image blur caused by the motion of the subject, camera shake (for mobile devices), and out-of-focus capture. (b) Faces in videos usually exhibit occlusion and a large range of pose variations.

in model ensembles is to train models separately for the holistic face image and for patches cropped around facial components. Model fusion is then performed offline at the feature or score level [Ding et al., 2016, Sun et al., 2014]. However, the accuracy promoted by model ensembles is at the cost of significantly increased time cost, which is impractical for VFR since each video usually contains dozens or even thousands of frames.

Here we approach the blur-robust representation learning problem from the perspective of training data. Since the volume of real-world video training data is small, we propose simulating large amounts of video frames from existing still face image databases. During training, we provide CNN with two training data streams: one composed of still face images, and the other composed of simulated video frames created by applying random artificial blur to the first stream. The network aims to classify each still image and its artificially blurred version into the same class; therefore, the learnt face representations must be blur-insensitive. To the best of our knowledge, this is the first CNN-based approach to solve the image blur problem in VFR.

To learn pose- and occlusion-robust representations for VFR efficiently, we propose a novel end-to-end ensemble CNN model called Trunk-Branch Ensemble CNN (TBE-CNN). TBE-CNN includes one trunk network and several branch networks. The trunk

network learns face representations for holistic face images, and each branch network learns representations for image patches cropped around one facial component. To speed up computation, the trunk and branch networks share the same low- and middle-level layers, while their high-level layers are optimized separately. This sharing strategy significantly reduces the computational cost of the ensemble model and at the same time exploits each model's uniqueness. The output feature maps by the trunk network and branch networks are fused by concatenation to form a comprehensive face representation.

Furthermore, to enhance TBE-CNN's discriminative power, we propose a Mean Distance Regularized Triplet Loss (MDR-TL) function to train TBE-CNN in an end-to-end fashion. Compared with the popular triplet loss function [Schroff et al., 2015], MDR-TL takes full advantage of label information and regularizes triplet loss by considering the global distribution of the training samples.

The efficacy of the proposed algorithm is systematically evaluated on three popular video face databases: PaSC [Beveridge et al., 2013], COX Face [Huang et al., 2015b], and YouTube Faces [Wolf et al., 2011a]. The evaluation is conducted for S2V, V2S, and V2V tasks. Extensive experiments on the three datasets indicate that the proposed algorithm achieves superior performance.

The remainder of the chapter is organized as follows. Section 5.2 briefly reviews related VFR works. The TBE-CNN model that handles image blur and pose variations for VFR is described in Section 5.3. The TBE-CNN training strategy is introduced in Section 5.4. Face matching using the proposed approach is illustrated in Section 5.5. Experimental results are presented in Section 5.6, leading to conclusions in Section 5.7.

5.2 Related Works

We review the literature in two parts: 1) video-based face recognition, and 2) deep learning methods for face recognition.

5.2.1 Video-based Face Recognition

Existing studies on VFR can be divided into three categories: (i) approaches for frame quality evaluation, (ii) approaches that exploit redundant information contained

between video frames, and (iii) approaches that attain robust feature extraction from each frame.

Frame quality evaluation is mainly utilized for key frame selection from video clips, such that only a subset of best-quality frames is selected for efficient face recognition. For a systematic summary of frame quality evaluation methods, we direct readers to two recent works [Mau et al., 2013, Phillips et al., 2013].

In contrast to frame quality evaluation methods, a number of recent VFR studies attempt to make use of redundant information contained between video frames. Algorithms that fall in this category include sequence-based methods, dictionary-based methods, and image set-based methods [Barr et al., 2012]. The sequence-based methods aim to extract person-specific facial dynamics from continuous video frames [Bicego et al., 2006, Hadid and Pietikäinen, 2009], which means that they rely on robust face trackers. The dictionary-based methods construct redundant dictionaries using video frames and employ sparse representation-based classifiers for classification [Chen et al., 2012b, Liu et al., 2014]. Due to the large size of the constructed dictionaries, the dictionary-based methods are often inefficient. The image set-based methods model the distribution of video frames using various techniques, e.g., affine/convex hull [Hu et al., 2012, Zhu et al., 2014a], linear subspace [Huang et al., 2015b], and manifold methods [Cui et al., 2012, Harandi et al., 2011]. They then measure the between-distribution similarity to match two image sets. The downside of image set modeling is that it is sensitive to the variable volume of video frames and complex facial variations that exist in real-world scenarios [Shao et al., 2015].

Extracting high-quality face representations has always been a core task in face recognition [Ding and Tao, 2016, Wolf et al., 2011b]. In contrast to still face images, video frames usually suffer from severe image blur because of the relative motion between the subjects and the cameras. Two types of methods have been introduced to reduce the impact of image blur: deblur-based methods [Nishiyama et al., 2011] and blur-robust feature extraction-based methods [Chan et al., 2013b, Gopalan et al., 2012]. The former method first estimates a blur kernel from the blurred image and then deblurs the face image prior to feature extraction. However, the estimation of the blur kernel is challenging, as it is an ill-posed problem. Of the blur-robust feature extraction methods, Ahonen et al. [2008] proposed to employ the blur-insensitive Local Phase Quantization (LPQ) descriptor for facial feature extraction, which has

been widely used in VFR applications [Beveridge et al., 2015]. However, to the best of our knowledge, no CNN-based method has yet been used to handle the image blur problem in VFR. Furthermore, similar to still face images, faces in video frames exhibit rich pose, illumination, and expression variations and occlusion; therefore, existing studies tend to directly extend feature extractors designed for SIFR to VFR [Chen et al., 2016, Li et al., 2014a, Parkhi et al., 2014].

Our proposed approach falls into the third category of methods. Compared to previous VFR approaches, we propose an efficient CNN model to automatically learn face representations that are robust to image blur, pose variations, and occlusion.

5.2.2 Deep Learning Methods for Face Recognition

Representation learning with CNNs provides an effective tool for face recognition. Promising results have been obtained, but they are usually limited to SIFR [Parkhi et al., RoyChowdhury et al., 2015, Schroff et al., 2015]. For example, Taigman et al. [2014a] formulated face representation learning as a face identification problem in CNN. [Hu et al., 2014, Lu et al., 2015, Schroff et al., 2015, Sun et al., 2014] proposed deep metric learning methods to enhance the discriminative power of learnt face representations. In comparison, there are only limited studies on CNN-based VFR.

This is for a number of reasons. First, existing CNN-based approaches do not handle the peculiarities of video frames (such as severe image blur) very well. Training from real-world video data is an intuitive solution for VFR. However, existing face video databases are rather small, e.g., 2,742 videos in PaSC and 3,000 videos in COX Face. Second, the video data are highly redundant, because frames in the same video tend to be similar in image quality, expression, occlusion, and illumination conditions. Therefore, direct CNN training using real-world video data is prone to overfitting. As a result, the majority of existing works employ CNN models trained on large-scale still image databases for VFR [Parkhi et al., Taigman et al., 2014a] and ignore the difference in image quality between still images and video frames. Recently, Huang *et al.* [Beveridge et al., 2015] proposed pre-training CNN models with a large volume of still face images and then fine-tuning the CNN models with small real-world video databases. However, the fine-tuning strategy is suboptimal, as it only slightly adapts CNN parameters to the video data. In comparison, our proposed video data simulation

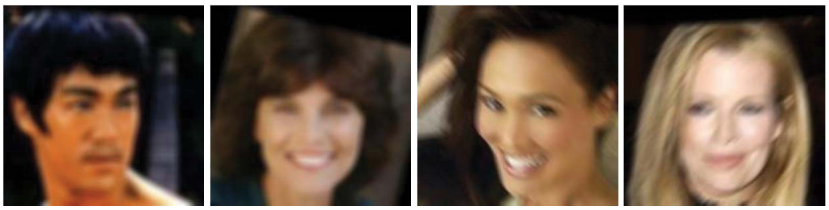
strategy enables direct optimization of all model parameters for VFR. Video data simulation can be regarded as a novel method for data augmentation. Compared to previous augmentation methods such as horizontal flipping and translation, video data simulation produces visually different images compared to the original still images, as illustrated in Fig. 5.2.

Ensemble models have been widely employed in SIFR to achieve robustness to pose variations and occlusion [Liu et al., 2015, Sun et al., 2014]. The principle is to train CNN models separately using the holistic face image and patches cropped around facial components. Model fusion is usually conducted offline by directly concatenating representations learnt by all models. Since image patches are less sensitive to pose variations and occlusion [Arashloo and Kittler, 2011], the ensemble system outperforms single models. However, there are drawbacks from two perspectives. First, the ensemble system is not an end-to-end system, which means its parameters are not simultaneously optimized for the VFR task. Second, extracting features separately from multiple models significantly reduces efficiency, which may be impractical for VFR since each video tends to have a number of frames to process. The Part-Stacked CNN (PS-CNN) model proposed in [Huang et al., 2015a] for fine-grained image classification is relevant to the current work. PS-CNN includes two models, one for the holistic image and one for object parts, and it dictates that object parts share all convolutional layers and a representation of each part is formed by cropping feature maps from the last convolutional layer. In comparison, TBE-CNN is more concise: it integrates the networks for the holistic image and all facial components into a single model by sharing their low- and middle-level layers. It is also more discriminative: the holistic image and each facial component have separate high-level convolutional layers to better exploit their complementary information.

5.3 Trunk-Branch Ensemble CNN for VFR

In this section, we make two contributions to VFR. First, we introduce a method for artificially simulating video training data for blur-robust face recognition. Second, the efficient TBE-CNN model for VFR is described.

(a)



(b)

Figure 5.2: Examples of the original still face images and simulated video frames. (a) original still images; (b) simulated video frames by applying artificial out-of-focus blur (the two figures on the left) and motion blur (the two figures on the right).

5.3.1 Artificially Simulated Video Data

As described above, most available video face databases are rather small and lack diversity in facial variations compared to still face image databases. We propose artificially generating video-like face data from existing large-scale still face image databases. Specifically, we simulate two challenges during surveillance or mobile camera imaging: motion blur and out-of-focus blur.

Due to face movement or mobile device camera shake during exposure, motion blur often appears in video frames. Supposing the relative motion is along a single direction during exposure, we can model the motion blur effect by one-dimensional local averaging of neighboring pixels using the following kernel:

$$k_m(i,j;L,\theta) = \begin{cases} \frac{1}{L}, & \text{if } \sqrt{i^2+j^2} \leq \frac{L}{2} \text{ and } \frac{i}{j} = -\tan\theta, \\ 0, & \text{otherwise}, \end{cases} \tag{5.1}$$

where (i,j) is the pixel coordinate originating from the central pixel, L is the size of the kernel and indicates the motion distance during exposure, and θ denotes the motion direction [Wang and Tao, 2014].

Due to the limited depth of field (DOF) of cameras and the large motion range of

faces in videos, out-of-focus blur also occurs in video frames and can be simulated using a uniform kernel [Wang and Tao, 2014] or a Gaussian kernel [Nishiyama et al., 2011]. In this chapter, we employ the Gaussian kernel in the following form:

$$k_o(i,j) = \begin{cases} C \cdot \exp\left(-\frac{i^2+j^2}{2\sigma^2}\right), & \text{if } i \leq \frac{R}{2} \text{ and } j \leq \frac{R}{2}, \\ 0, & \text{otherwise}, \end{cases} \tag{5.2}$$

where σ and R denote the magnitude and size of the kernel, respectively. C is a constant which ensures the kernel has a unit volume.

Given one still face image I_s and one blur kernel k, the simulated video frame I_v can be obtained by convolution:

$$I_v = I_s * k. \tag{5.3}$$

Based on the above description, we blur each still face image using one randomly sampled blur type from the following 38 possibilities. Specifically, we choose the value of L from $\{7, 9, 11\}$ and the value of θ from $\{0, \pi/4, \pi/2, 3\pi/4\}$; therefore, there are 12 motion blur choices. Similarly, we set the value of R as 9 and randomly choose the value of σ from $\{1.5, 3.0\}$; i.e., there are 2 choices for out-of-focus blur. We also enrich the blur types by sequentially conducting out-of-focus blur and motion blur, of which there are 24 combinations. Samples of original still images and the corresponding simulated video frames are illustrated in Fig. 5.2. Since we obtain one blurred image from each still image, we obtain two training data streams of equal size, i.e., one stream composed of the original still images and the other stream composed of the same number of blurred images. For CNN training, we provide both training data streams to CNN simultaneously. Since we encourage each still image and its blurred version to be classified into the same class, CNN automatically learns blur-robust face representations.

5.3.2 Trunk-Branch Ensemble CNN

We propose the Trunk-Branch Ensemble CNN (TBE-CNN) model to efficiently learn pose- and occlusion-robust face representations. TBE-CNN incorporates one trunk network and several branch networks. The trunk network is trained to learn face

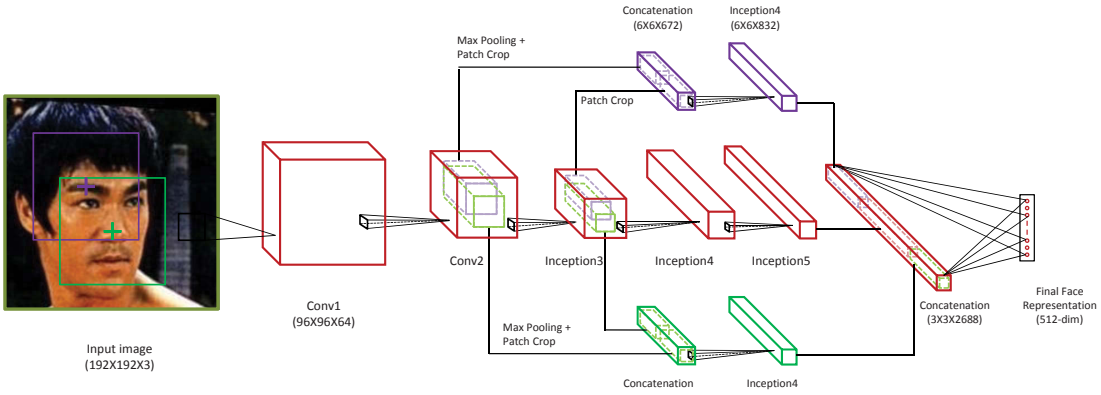


Figure 5.3: Model architecture for Trunk-Branch Ensemble CNN (TBE-CNN). Note that a max pooling layer is omitted for simplicity following each convolution module, e.g., Conv1 and Inception 3. TBE-CNN is composed of one trunk network that learns representations for holistic face images and two branch networks that learn representations for image patches cropped around facial components. The trunk network and the branch networks share the same low- and middle-level layers, and they have individual high-level layers. The output feature maps of the trunk network and branch networks are fused by concatenation. The output of the last fully connected layer is utilized as the final face representation of one video frame.

representations for holistic face images, and each branch network is trained to learn face representations for image patches cropped from one facial component. In this chapter, the trunk network implementation is based on GoogLeNet [Szegedy et al., 2015], the most important parameters of which are tabulated in Table 5.1. For the other model parameters, we directly follow its original configuration [Szegedy et al., 2015]. As shown in Table 5.1, we divide the GoogLeNet layers into three levels: the low-level layers, middle-level layers, and high-level layers. The three layer levels successively extract features from the low to the high-level. Since low- and middle-level features represent local information, the trunk network and branch networks can share low- and middle-level layers. In comparison, high-level features represent abstract and global information; therefore, different models should have separate high-level layers. Based on the above observations, the TBE-CNN architecture is illustrated in Fig. 5.3.

The trunk network extracts features from raw pixels. On top of Inception 5 of the trunk network, we reduce the size of its feature maps to $3\times3\times1024$ by max pooling. For each branch network, we directly crop feature maps from Conv2 and Inception 3 module outputs of the trunk network instead of computing low- and middle-level

Table 5.1: Trunk Network Parameters (GoogLeNet)

	Type (Name)	Kernel Size/ Stride	Output Size	Depth
Low-Level	convolution (Conv1)	7×7/2	96×96×64	1
	max pool	2×2/2	48×48×64	0
	convolution (Conv2)	3×3/1	48×48×192	2
	max pool	2×2/2	24×24×192	0
Middle-L	inception (3a)	-	24×24×256	2
	inception (3b)	-	24×24×480	2
	max pool	2×2/2	12×12×480	0
High-Level	inception (4a)	-	12×12×512	2
	inception (4b)	-	12×12×512	2
	inception (4c)	-	12×12×512	2
	inception (4d)	-	12×12×528	2
	inception (4e)	-	12×12×832	2
	max pool	2×2/2	6×6×832	0
	inception (5a)	-	6×6×832	2
	inception (5b)	-	6×6×1024	2
	max pool	2×2/2	3×3×1024	1
	Dropout (0.4)	-	3×3×1024	1
	Fully-connected	-	512	1

features from scratch. The cropping size and position are propositional to those of the patches of interest from the input image, as illustrated in Fig. 5.3. The size of the feature maps cropped from the Conv2 output is reduced by a half by max pooling and then concatenated with the feature maps cropped from the Inception 3 output. The concatenated feature maps form the input of the branch network. To promote efficiency, each branch network includes only one Inception 4 module. Similar to the trunk network, the size of feature maps of the branch network is reduced to 3×3×832 by max pooling.

We include two branch networks in TBE-CNN, as illustrated in Fig. 5.3. The output

feature maps of the trunk network and branch networks are fused by concatenation to form an over-complete face representation, whose dimension is reduced to 512 by one fully connected layer. The 512-dimensional feature vector is utilized as the final face representation of one video frame.

5.4 TBE-CNN Training

We propose a stage-wise training strategy to effectively optimize the TBE-CNN model parameters. In the first stage, the trunk network illustrated in Table 5.1 is trained alone using softmax loss as the penalty. In the second stage, the trunk network parameters are fixed, and each of the branch networks is trained with softmax loss[1]. After the trunk network and all branch networks are pre-trained, we fine-tune the complete model illustrated in Fig. 5.3 with softmax loss to fuse the trunk and branch networks. Finally, to enhance the discriminative power of learnt face representations, we propose a novel deep metric learning method called Mean Distance Regularized Triplet Loss (MDR-TL) to fine-tune the complete network.

The softmax loss has the advantage of convergence efficiency, but the penalty is imposed on the classification accuracy on training data rather than the discriminative power of learnt face representations. In comparison, the convergence rate of MDR-TL or triplet loss is slower, but they directly optimize the discriminative power of the face representations. We next introduce the details of MDR-TL.

5.4.1 Mean Distance Regularized Triplet Loss

Existing deep metric learning methods for face recognition include pairwise loss [Sun et al., 2014] and triplet loss [Schroff et al., 2015]. Both methods rely on sampling effective image pairs or triplets from all possibilities. However, since the optimization is based on each individual image pair or triplet, the global distribution of training samples is neglected, which has a negative impact on face recognition. For example, the training samples in Fig. 5.4(a) satisfy the triplet constraint; however, it is difficult to find an ideal threshold for face verification due to nonuniform intra-class and inter-

[1] We include one 256-dimensional fully connected layer after the Inception 4 module of each branch network when training it alone.

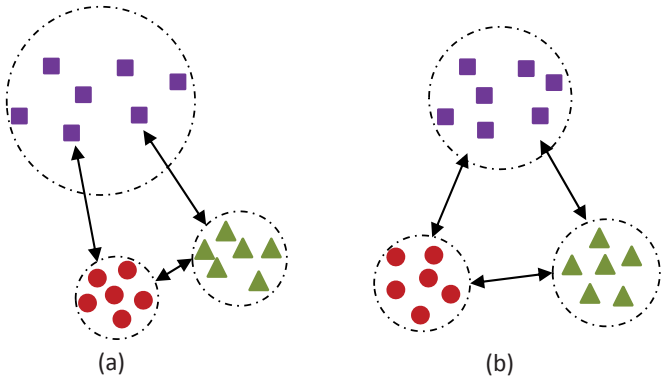


Figure 5.4: The principle of Mean Distance Regularized Triplet Loss (MDR-TL). (a) Triplets sampled in the training batch satisfy the triplet constraint (Eq. 5.4). However, due to the non-uniform intra-class and inter-class sample distributions, it is hard to select an ideal threshold for face verification. (b) MDR-TL regularizes triplet loss by setting a margin for the distance between subject mean representations so that samples of different subjects are uniformly distributed.

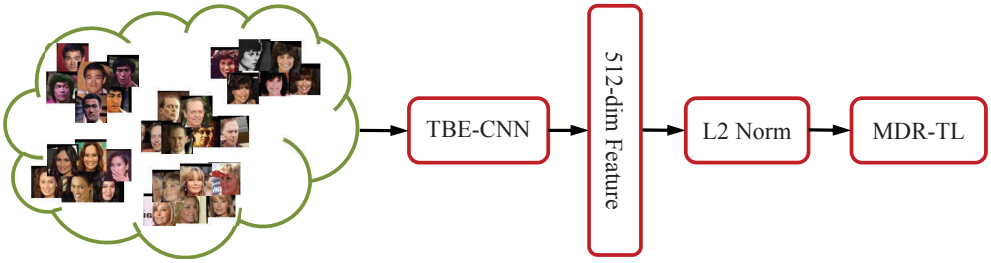


Figure 5.5: Illustration of TBE-CNN training with MDR-TL. MDR-TL is employed to further enhance the discriminative power of learnt face representations.

class sample distributions. To overcome this problem, we propose the MDR-TL loss function, which regularizes the triplet loss by taking the global distribution of training samples into consideration.

As illustrated in Fig. 5.5, the 512-dimensional face representation extracted by TBE-CNN is ℓ_2-normalized as the MDR-TL input. We denote the ℓ_2-normalized face representation of one image x as $f(x) \in \mathbb{R}^d$. Two constraints are included in MDR-TL. In the first, triplets sampled in a training batch should satisfy the triplet constraint [Schroff et al., 2015]:

$$\|f(x^a) - f(x^p)\|_2^2 + \beta < \|f(x^a) - f(x^n)\|_2^2, \tag{5.4}$$

where $f(x^a)$, $f(x^p)$, and $f(x^n)$ represent the ℓ_2-normalized face representations of the anchor point, positive point, and negative point, respectively. In the second, the mean representations of different subjects should be well separated to ensure that samples of different subjects are uniformly distributed. We realize this constraint by enforcing a margin α between the mean representation $\hat{\mu}_c$ of one subject and its nearest mean representation $\hat{\mu}_c^n$:

$$\|\hat{\mu}_c - \hat{\mu}_c^n\|_2^2 > \alpha, \tag{5.5}$$

where

$$\hat{\mu}_c = \frac{\mu_c}{\|\mu_c\|}, \tag{5.6}$$

and

$$\mu_c = \frac{1}{N_c} \sum_{i=1}^{N_c} f(x_{ci}). \tag{5.7}$$

N_c is the number of images of the cth subject and $f(x_{ci})$ denotes the face representation of the ith image for the cth subject. The nearest mean representation $\hat{\mu}_c^n$ for $\hat{\mu}_c$ is detected online within the same batch.

Based on the above analysis, we formulate MDR-TL as the following optimization problem:

$$\min_f L(f) = L_{triplet}(f) + L_{mean}(f), \tag{5.8}$$

where $L_{triplet}(f) =$

$$\frac{1}{2N} \sum_{i=1}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \beta \right]_+, \tag{5.9}$$

and

$$L_{mean}(f) = \frac{1}{2P} \sum_{c=1}^{C} max(0, \alpha - \|\hat{\mu}_c - \hat{\mu}_c^n\|_2^2). \tag{5.10}$$

In Eq. 5.8, we assume equal weights for $L_{triplet}(f)$ and $L_{mean}(f)$, which empirically works well. N is the number of triplets that violate the constraint in Eq. 5.4. β is the

margin for the triplet constraint, C is the number of subjects in the current batch, and P is the number of mean representations that violate the constraint in Eq. 5.5.

We optimize Eq. 5.8 using the standard stochastic gradient descent with momentum [Jia et al., 2014]. The gradient of L with respect to $f(x_{ci})$ is derived as follows,

$$\frac{\partial L}{\partial f(x_{ci})} = \frac{\partial L_{triplet}}{\partial f(x_{ci})} + \frac{\partial L_{mean}}{\partial f(x_{ci})}, \tag{5.11}$$

where

$$\frac{\partial L_{mean}}{\partial f(x_{ci})} = -\frac{1}{P}(\sum_{j=1,j\neq c}^{C} w_j(\hat{\mu}_c - \hat{\mu}_j)^T)\frac{\partial \hat{\mu}_c}{\partial f(x_{ci})}. \tag{5.12}$$

w_j equals 1 if the constraint in Eq. 5.5 is violated, and $\hat{\mu}_j$ is the nearest neighbor of $\hat{\mu}_c$ or $\hat{\mu}_c$ is the nearest neighbor of $\hat{\mu}_j$. Otherwise, w_j equals 0.

$$\begin{aligned}\frac{\partial \hat{\mu}_c}{\partial f(x_{ci})} &= \|\mu_c\|^{-1}\frac{\partial \mu_c}{\partial f(x_{ci})} + \mu_c\frac{\partial \|\mu_c\|^{-1}}{\partial f(x_{ci})} \\ &= \frac{1}{N_c\|\mu_c\|}(I - \hat{\mu}_c\hat{\mu}_c^T).\end{aligned} \tag{5.13}$$

The derivative of L with respect to model parameters can be computed via the chain rule. As illustrated in Fig. 5.4(b), by setting a margin between the mean representations of different subjects, the triplet loss is regularized by the global distribution of training samples. This regularization is empirically verified as important below.

5.5 VFR with TBE-CNN

We next employ the TBE-CNN model for VFR. Given a video, we first augment its video frames by horizontal flipping. Then, all video frames pass through the TBE-CNN network, and fusing their outputs by average pooling produces the compact video representation. The representation of one still image can be extracted in a similar way by assuming it is a single-frame video. For all three settings - S2V, V2S, and V2V

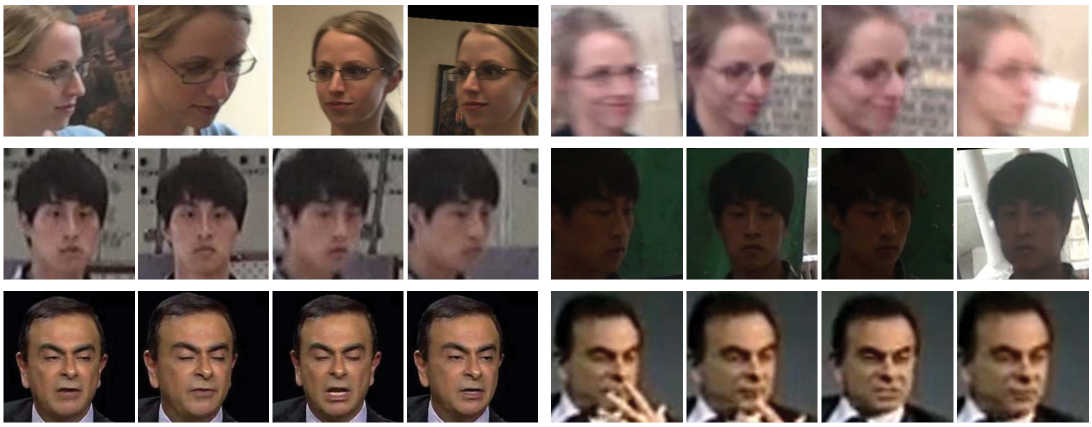


Figure 5.6: Sample video frames after normalization: PaSC (first row), COX Face (second row), and YouTube Faces (third row). For each database, the four frames on the left are sampled from a video recorded under relatively good conditions, and the four frames on the right are selected from low-quality video.

matching - we consistently employ the cosine metric to calculate the similarity between the two representations y_1 and y_2:

$$Sim(y_1, y_2) = \frac{y_1^T y_2}{\|y_1\|\|y_2\|}. \tag{5.14}$$

5.6 Experiments

We now systematically evaluate the proposed TBE-CNN framework for VFR. Experiments are conducted on three publicly available large-scale video face databases: PaSC, COX Face, and YouTube Faces. Example images are shown in Fig. 5.6. In particular, we emphasize experiments on the first two databases, since their data well simulates real-world surveillance videos.

The PaSC database [Beveridge et al., 2013] contains 2,802 videos of 265 subjects. The videos were recorded by multiple sensors in varied indoor and outdoor locations. They are divided into two sets: a control set and a handheld set. High-end cameras installed on tripods captured videos in the control set to produce images of good quality, as illustrated by the first four images in the first row in Fig. 5.6. Videos in the handheld set were captured by five handheld video cameras. For each video,

the subject was asked to carry out actions to create a wide range of poses at variable distances from the camera. Faces in the video exhibit serious motion and out-of-focus blur and rich pose variations. Here, we adopted the officially defined V2V matching protocol for face verification [Beveridge et al., 2013].

The COX Face database [Huang et al., 2015b] incorporates 1,000 still images and 3,000 videos of 1,000 subjects. A high-quality camera in well-controlled conditions captured still images to simulate ID photos. The videos were taken while the subjects were walking in a large gym to simulate surveillance. Three cameras at different locations were installed to capture videos of the walking subject simultaneously. Videos captured by the three cameras created three subsets: Cam1, Cam2, and Cam3. Since COX Face has significantly more subjects compared to PaSC, it is an ideal database for face identification. The standard S2V, V2S, and V2V matching protocols [Huang et al., 2015b] for face identification were adopted.

The YouTube Faces database [Wolf et al., 2011a] includes 3,425 videos of 1,595 subjects. All videos were downloaded from the YouTube website. Since the majority of subjects in this database were in interviews, there is no obvious image blur or pose variation. Instead, this database is low resolution and contains serious compression artifacts, as shown in the last four video frames of the third row in Fig. 5.6. 5,000 video pairs were collected from the videos, which were divided into 10 equally sized splits. Each split contained 250 homogeneous video pairs and 250 heterogeneous video pairs. Following the "restricted" protocol defined in [Wolf et al., 2011a], we report the mean verification accuracy and the standard error of the mean (S_E) on the 10 splits.

For all three video databases, we employ the face detection results provided by the respective databases. There are 60 videos in PaSC that have no face being detected. We ignore these 60 videos since their similarity score to any video is NaN, and thus have no impact on the verification rate or Receiver Operating Characteristic (ROC) curve. Each detected face is normalized and resized to 192×192 pixels using an affine transformation based on the five facial feature points detected by [Sun et al., 2013a]. Sample face images after normalization are shown in Fig. 5.6. A number of experiments are conducted. First, the effectiveness of video training data simulation for CNN is tested. Second, the MDR-TL performance is evaluated. Third, the trunk network performance and TBE-CNN are compared. Finally, the performance of the complete TBE-CNN framework is compared to state-of-the-art VFR methods on the

PaSC, COX Face, and YouTube Faces databases.

5.6.1 Implementation Details of TBE-CNN

The proposed TBE-CNN model is trained on the publicly available CASIA-WebFace database [Yi et al., 2014] and deployed in experiments on the three video face databases described above. The CASIA-WebFace database contains 494,414 images of 10,575 subjects. We augment the CASIA-WebFace database by horizontal flipping and image jittering. After augmentation, the size of the training set is about 2.68 million, which forms the still image training data stream. According to the description in Section 5.3, we blur each still image using a randomly chosen kernel to form the simulated video training data stream.

During training, the batch size is set to 90 for both the still image stream and simulated video stream to produce a complete batch size of 180. Following the stage-wise training framework introduced in Section 5.4, the trunk network is trained for 13 epochs using softmax loss, with the learning rate gradually decreased from 0.01 to 0.001. Then, the branch networks are trained for 4 epochs. Next, the complete TBE-CNN model is fine-tuned for another 4 epochs with softmax loss to fuse the trunk and branch networks with a small learning rate of 0.001. Finally, the TBE-CNN model is fine-tuned with MDR-TL for one more epoch with the learning rate of 0.001. The open-source deep learning package Caffe [Jia et al., 2014] is utilized to train the deep models.

The first three experiments are conducted on PaSC. For these three experiments, we manually remove the falsely detected faces from the face detection results provided by the database[1] to help us to accurately reflect the effectiveness of proposed algorithms. For the fourth experiment, we provide results based on both the original face detections and manually corrected face detections.

5.6.2 Effectiveness of Simulated Video Training Data

This experiment provides evidence to justify the proposed video data simulation strategy. The evaluation is based on the trunk network with softmax loss. Four types

[1]The defect in face detection is because several different faces appeared in the same video. The list of manually removed faces is available upon request.

of training data are evaluated: the still image training data stream alone (SI), the simulated video training data stream alone (SV), the fine-tuning (FT) strategy adopted for VFR [Beveridge et al., 2015], and the two-stream training data (TS) that combines both SI and SV. For SI, SV, and TS, the networks are trained with the same number of iterations (13 epochs), and the amount of augmented training data is 2.6 million, 2.6 million, and 5.2 million, respectively. For FT, we fine-tune the network produced by SI with real-world video data. Following [Beveridge et al., 2015], we use all video frames in the COX Face database and the PaSC training set for fine-tuning. The two video sets incorporate 0.45 million video frames, comparable in volume to the CASIA-WebFace database. For fair comparison with TS, we augment the video frames with horizontal flipping and jittering to 2.6 million. Therefore, the total amount of training data for FT is 5.2 million. The learning rate of the fine-tuning stage for FT is set to 0.001, and we observe that its performance saturates on PaSC after one epoch.

The verification rates at 1% False Acceptance Rate (FAR) of the above four types of training data are tabulated in Table 5.2, and the corresponding ROC curves are illustrated in Fig. 5.7. While SI slightly outperforms SV on the control set, SV significantly outperforms SI by as much as 5.6% on the handheld set. This is due to the difference in image quality between the control and handheld sets. Specifically, high-end cameras recorded video frames in the control set; therefore, the image quality is similar to that of the training data for SI. In comparison, low-quality mobile cameras captured videos in the handheld set and its frames sufferred from severe image blur, as illustrated in Fig. 5.6. The clear performance advantage of SV on the handheld set justifies the use of simulated video training data.

The performance of FT is comparable to SV, but at the cost of more training data and longer training time. In comparison, TS significantly outperforms SI, SV, and FT on both sets. Compared to SI and SV, TS combines the training data of both still images and simulated video frames, which regularizes the model to learn blur-robust representations. Compared to FT, which has the same size of training data, TS shows a performance advantage of 5.9% and 5.6% for the control and handheld sets, respectively. This is for two main reasons: first, there is only slight adaptation of CNN model parameters by FT to the VFR task compared to TS, which directly optimizes all model parameters of CNN for VFR; second, frames in the same video are highly redundant. Therefore, fine-tuning with a small amount of real-world video data may

Table 5.2: Verification Rates (%) at 1% FAR on PaSC with Different Types of Training Data

	# Training Data (Augmented)	Control Set	Handheld Set
SI	2.6M	83.72	68.20
SV	2.6M	81.89	73.80
FT	5.2M	81.41	73.70
TS	5.2M	**87.31**	**79.33**

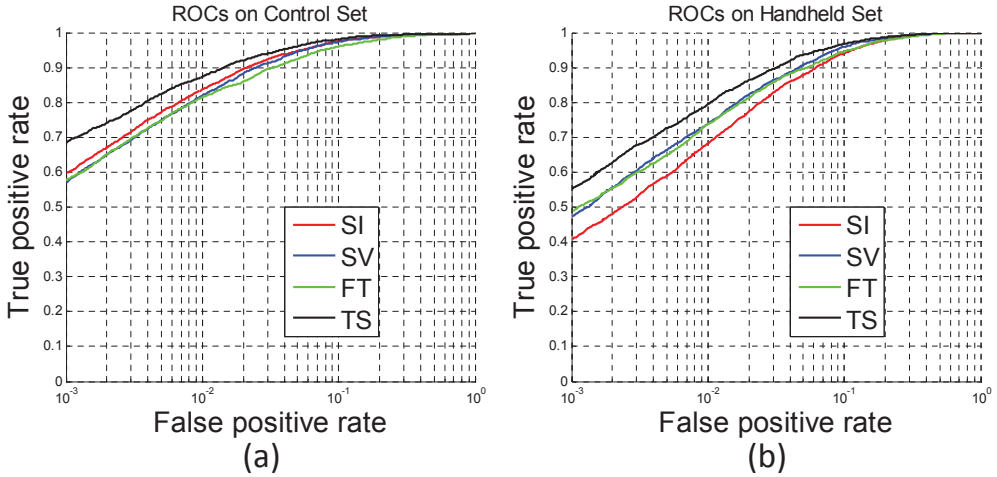


Figure 5.7: ROC curves of the trunk network trained with different types of training data on the PaSC database. (a) Comparison on the control set; (b) comparison on the handheld set.

suffer from over-fitting.

5.6.3 Effectiveness of MDR-TL

This experiment evaluates the performance of the proposed MDR-TL loss function. Two baseline loss functions for deep metric learning are compared: pairwise loss [Zhang et al., 2009] and triplet loss [Schroff et al., 2015]. Similar to the previous experiment, performance comparisons are based on the trunk network. Two representative training data types are considered: 1) SI, which is the most widely used training data type; and 2) our proposed TS. The models produced in the previous

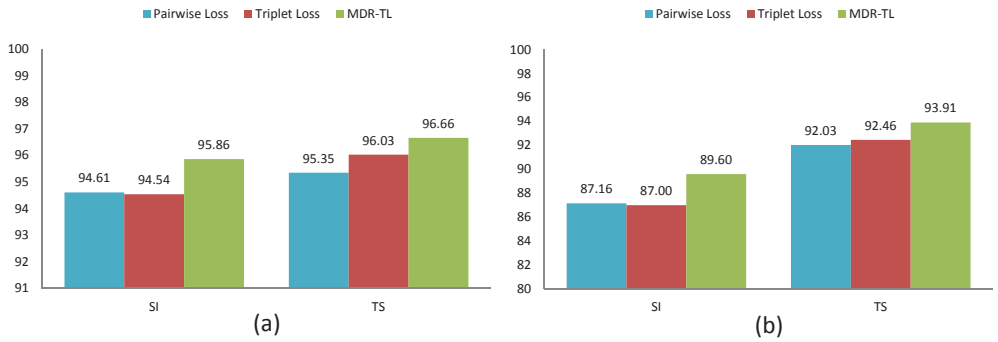


Figure 5.8: Verification rates at 1% FAR with different loss functions on the PaSC database. SI and TS stand for two representative types of training data. (a) Comparison on the control set; (b) comparison on the handheld set.

experiment are fine-tuned with a learning rate of 0.001 for one epoch with each of the three loss functions. We observe that the performance saturates after one epoch of fine-tuning. Image pairs or triplets are sampled within each batch online. To ensure that sufficient image pairs or triplets are sampled, we re-arrange the training samples such that each subject in a batch contains 6 images. Therefore, there are 30 subjects in total for a batch size of 180. For the pairwise loss, our implementation is based on the Discriminative Locality Alignment (DLA) model [Zhang et al., 2009], which has three parameters to tune. We empirically find that sampling all positive pairs and the same number of hard-negative pairs is optimal for DLA. For the triplet loss, following the specifications in [Schroff et al., 2015], we utilize all positive image pairs and randomly sample semi-hard negative samples to compose triplets. We traverse the value of the margin parameter β within $\{0.1, 0.2, 0.3, 0.4\}$ and report its best performance. Similarly, we traverse the value of the margin parameter α for MDR-TL within $\{1.6, 1.8, 2.0, 2.2\}$ and report the best result. It is worth noting that our implementation of the $L_{triplet}(f)$ term in Eq. 5.8 for both triplet loss and MDR-TL is exactly the same. The optimal values for α and β are 2.0 and 0.2, respectively. The verification rates at 1% FAR of the three loss functions are compared in Fig. 5.8.

It is shown that the performance of pairwise loss (DLA) is comparable to triplet loss. However, the major disadvantage of DLA is that it has more parameters; therefore, it is more difficult to tune in practice. MDR-TL outperforms the baseline loss functions consistently under all settings. In particular, for the most challenging setting, i.e., the handheld set of PaSC with SI training data, MDR-TL outperforms the

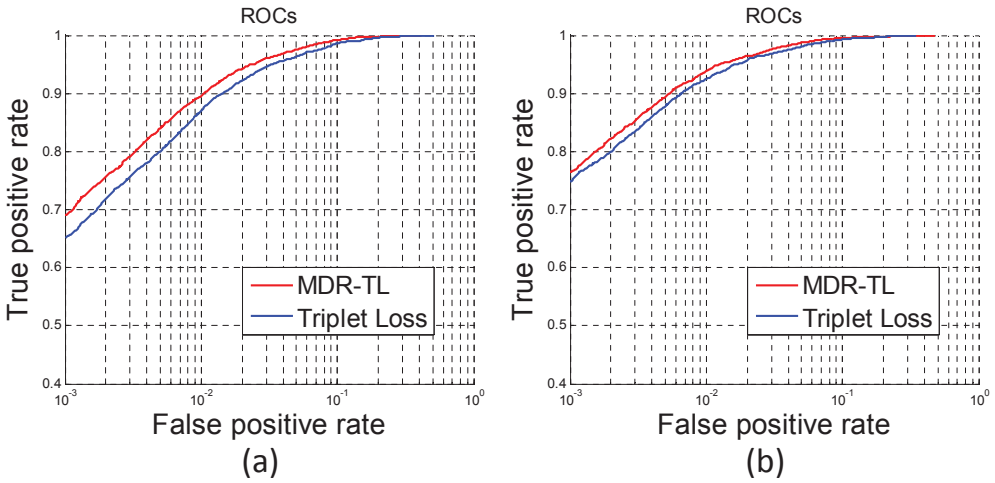


Figure 5.9: ROC curves of MDR-TL and triplet loss functions on the handheld set of PaSC. (a) SI training data; (b) TS training data.

triplet loss by as much as 2.6%. The ROC curves of the two loss functions on the handheld set of PaSC are plotted in Fig. 5.9. The experimental results justify the role of $L_{mean}(f)$ in Eq. 5.8 as an effective regularization term to triplet loss.

The comparisons in Fig. 5.8 and Fig. 5.9 also convincingly show that the proposed TS strategy to compose CNN training data is essential to achieve blur-robustness in VFR. Briefly, for the control and handheld sets of PaSC, CNN performance with TS training data is higher than that with SI training data by around 1.0% and 5.0%, respectively.

5.6.4 Effectiveness of Trunk-Branch Fusion

The verification rates at 1% FAR of TBE-CNN and trunk network with TS training data on PaSC are compared in Fig. 5.10. Comparison is based on the softmax loss. Two results are presented, i.e., with and without the batch normalization (BN) [Ioffe and Szegedy, 2015] layers employed. To enable the training of TBE-CNN by a single GPU, we only add BN layers after convolutional layers of the Inception 4 and Inception 5 modules as BN consumes video memory. The performance of TBE-CNN is considerably higher than the trunk network. In particular, on the more challenging handheld set, the margin is as large as 5.82% and 3.48% without and with BN layers, respectively. The ROC curves of the two models on the handheld set are plotted

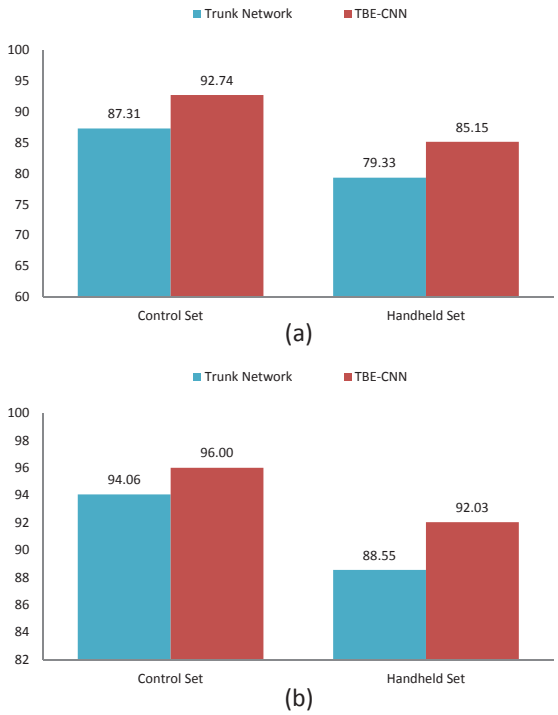


Figure 5.10: Verification rates (%) at 1% FAR by the trunk network and TBE-CNN. Comparison is based on the softmax loss. (a) Performance comparison without BN layers; (b) performance comparison with BN layers.

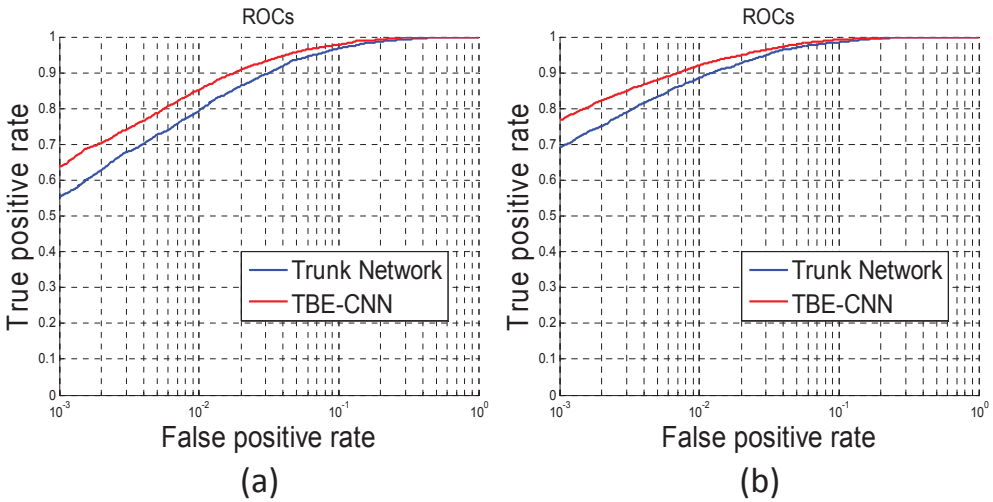


Figure 5.11: ROC curves of the trunk network and TBE-CNN on the handheld set of PaSC. (a) Without BN layers; (b) with BN layers.

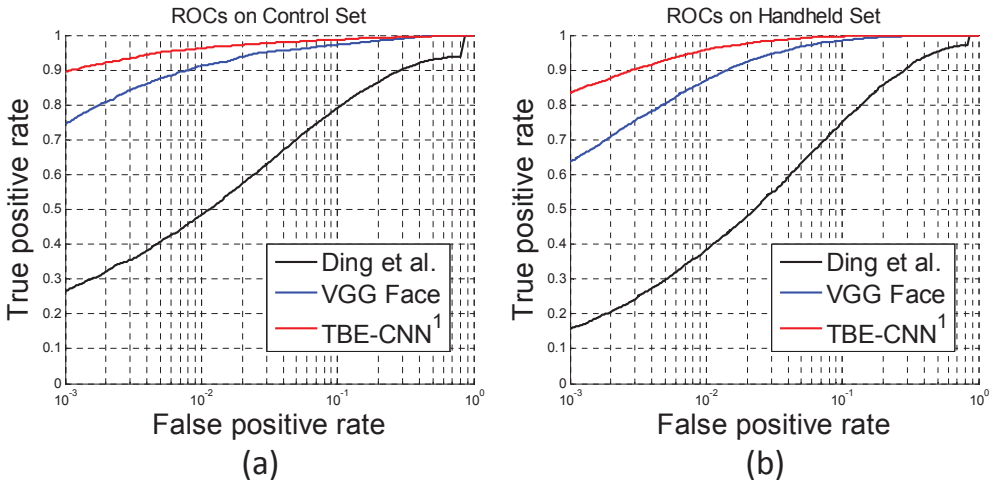


Figure 5.12: ROC curves of TBE-CNN and state-of-the-art methods on the PaSC control and handheld sets. The original face detection results from the database are employed for all methods. (a) Control set; (b) handheld set.

in Fig. 5.11. The comparisons suggest that TBE-CNN effectively makes use of the complementary information between the holistic face image and facial components. It is important to note that TBE-CNN is quite efficient in both time and video memory costs. In brief, the time and memory costs of TBE-CNN are only 1.26× and 1.2× those of the trunk network, respectively. In comparison, if there were no layer sharing, the total time and memory costs of the three networks would be 1.71× and 1.6× those of the trunk network, respectively.

5.6.5 Performance Comparison on PaSC

We next present the best performance of TBE-CNN on PaSC with TS training data and MDR-TL fine-tuning. For fair comparison with existing approaches, we present both results of TBE-CNN with and without BN layers. Similar to the previous experiment, we only add BN layers after convolutional layers of the Inception 4 and Inception 5 modules to save video memory. Performance comparison is illustrated in Fig. 5.12 and Table 5.3.

The performance of the VGG Face model [Parkhi et al.] is reported according to the model published by the authors. According to the description in [Parkhi et al.], we fine-tune the last fully connected layer of the VGG Face model with triplet loss

Table 5.3: Verification Rates (%) at 1% FAR of Different Methods on PaSC

	# Training Data (Original)	Control Set	Handheld Set
Ding *et al.* [Beveridge et al., 2015]	0.01M	48.00	38.00
Huang *et al.* [Beveridge et al., 2015]	0.5M	58.00	59.00
HERML [Huang et al., 2015c]	0.5M	46.61	46.23
VGG Face [Parkhi et al.]	2.62M	91.25	87.03
TBE-CNN[1]	0.49M	**95.83**	**94.80**
TBE-CNN[1]**+BN**	0.49M	**96.23**	**95.85**
TBE-CNN[2]**+BN**	0.49M	**97.80**	**96.12**

on the CASIA-WebFace database. For VGG Face model testing, we employ the same strategy as TBE-CNN to extract video representations, as described in Section 5.5.

In Table 5.3, TBE-CNN[1] denotes the TBE-CNN performance with the original face detection results from the database. TBE-CNN[2] denotes the performance based on manually cleaned face detection results. It is clear that the proposed approach achieves the best performance on both the control and handheld sets. In particular, the proposed approach outperforms the VGG Face model [Parkhi et al.] by 4.98% and 8.82% on the control and handheld sets, respectively. It is worth noting that the VGG Face model is trained with over 2.6 million still image data and aggressive data augmentation. In comparison, the size of our original training data is only 0.49 million. With more training data, we believe that TBE-CNN can perform even better.

Moreover, TBE-CNN wins the first place with a considerable margin in the BTAS 2016 Video Person Recognition Evaluation, which is also based on the PaSC database [BTA]. With fully automatic face detection, alignment, and recognition, our four-model ensemble system achieves 98.0% and 97.0% verification rates at 1% FAR on the control and handheld sets of PaSC, respectively. The detailed comparisons on PaSC justify the effectiveness of the proposed methods for VFR.

Table 5.4: Rank-1 Identification Rates (%) under the V2S/S2V Settings for Different Methods on the COX Face Database

	V2S Identification Rate			S2V Identification Rate		
	V1-S	V2-S	V3-S	S-V1	S-V2	S-V3
PSCL [Huang et al., 2015b]	38.60 ± 1.39	33.20 ± 1.77	53.26 ± 0.80	36.39 ± 1.61	30.87 ± 1.77	50.96 ± 1.44
LERM [Huang et al., 2014]	45.71 ± 2.05	42.80 ± 1.86	58.37 ± 3.31	49.07 ± 1.53	44.16 ± 0.94	63.83 ± 1.58
VGG Face [Parkhi et al.]	88.36 ± 1.02	80.46 ± 0.76	90.93 ± 1.02	69.61 ± 1.46	68.11 ± 0.91	76.01 ± 0.71
TBE-CNN	**93.57 ± 0.65**	**93.69 ± 0.51**	**98.96 ± 0.17**	**88.24 ± 0.40**	**87.86 ± 0.85**	**95.74 ± 0.67**

Table 5.5: Rank-1 Identification Rates (%) under the V2V Setting for Different Methods on the COX Face Database

	V2-V1	V3-V1	V3-V2	V1-V2	V1-V3	V2-V3
PSCL [Huang et al., 2015b]	57.70 ± 1.40	73.17 ± 1.44	67.70 ± 1.70	62.77 ± 1.02	78.26 ± 0.97	68.91 ± 2.28
LERM [Huang et al., 2014]	65.94 ± 1.97	78.24 ± 1.32	70.67 ± 1.88	64.44 ± 1.55	80.53 ± 1.36	72.96 ± 1.99
HERML [Huang et al., 2015c]	95.10 ± -	96.30 ± -	94.20 ± -	92.30 ± -	95.40 ± -	94.50 ± -
VGG Face [Parkhi et al.]	94.51 ± 0.47	95.34 ± 0.32	96.39 ± 0.42	93.39 ± 0.56	96.10 ± 0.27	96.60 ± 0.52
TBE-CNN	**98.07 ± 0.32**	**98.16 ± 0.23**	**97.93 ± 0.20**	**97.20 ± 0.26**	**99.30 ± 0.16**	**99.33 ± 0.19**

5.6.6 Performance Comparison on COX Face

The rank-1 identification rates for TBE-CNN and state-of-the-art algorithms on the COX Face database are tabulated in Tables 5.4 and 5.5. In Table 5.4, Vi-S (S-Vi) represents the test using the i-th video set as probe (gallery) and the still images as the gallery (probe). Similarly, in Table 5.5, Vi-Vj represents the test using the i-th video set as probe and the j-th video set as the gallery. For both TBE-CNN and VGG Face [Parkhi et al.], we directly deploy the models in the previous experiment to evaluate their performance on COX Face.

Under all experimental settings, TBE-CNN achieves the best performance. It is also clear that the S2V and V2S tasks are significantly more difficult than the V2V task, suggesting a huge difference in distribution between the still image domain and video data domain. Interestingly, TBE-CNN has overwhelming performance advantage in S2V and V2S tasks. This may be because it is trained with both still image and simulated video data, meaning it can learn invariant face representations from still images and video data from the same subject. This is a great advantage for many real-world VFR tasks, e.g., watch-lists in video surveillance applications, where the gallery is composed of high-quality ID photos and the probe includes video frames captured by surveillance cameras. Furthermore, compared to image-set model-based VFR methods [Huang et al., 2014, 2015b,c], TBE-CNN has two key advantages: first, the extracted video representation is very compact, which means it is efficient for retrieval tasks; and second, the representation is robust to fluctuations in the volume of frames in a video.

5.6.7 Performance Comparison on YouTube Faces

Finally, we compare the face verification performance of TBE-CNN with state-of-the-art approaches on the YouTube Faces database. As mentioned above, professional photographers rather than surveillance video cameras usually recorded videos in this database. Therefore, they are free from image blur. Also, since the majority of subjects in the video were in interviews, there is little pose variation, as illustrated in Fig. 5.6. Instead, the video frames are low resolution and contain serious compression artifacts. These characteristics make experiments on the YouTube Faces database more similar to the traditional SIFR task rather than the real-world VFR task. Therefore, existing

Table 5.6: Mean Verification Accuracy on the YouTube Faces Database (Restricted Protocol)

	# Crop	# Training Data (Original)	Accuracy(%) $\pm$ S_E
Deep Face* [Taigman et al., 2014a]	1	4.4M	91.40 $\pm$ 1.10
DeepID 2+ [Sun et al., 2015]	50	0.29M	93.20 $\pm$ 0.20
FaceNet [Schroff et al., 2015]	2	260M	95.12 $\pm$ 0.39
VGG Face* [Parkhi et al.]	30	2.62M	91.60 $\pm$ -
TBE-CNN*	2	0.49M	93.84 $\pm$ 0.32
TBE-CNN	2	0.49M	**94.96 $\pm$ 0.31**

SIFR models that are trained with large-scale still image databases, e.g., the FaceNet model [Schroff et al., 2015], may be expected to show advantages in this experiment. The mean verification rates under the "restricted" protocol [Wolf et al., 2011a] of state-of-the-art approaches are tabulated in Table 5.6. Corresponding ROC curves are plotted in Fig. 5.13. Under this protocol, the subject identity labels of YouTube Faces database are not allowed to be used for training. In Table 5.6, * denotes the models without training or fine-tuning by deep metric learning methods.

TBE-CNN outperforms previous approaches such as Deep Face [Taigman et al., 2014a], DeepID 2+ [Sun et al., 2015], and VGG Face [Parkhi et al.]. Its performance is comparable to the FaceNet model [Schroff et al., 2015] trained on as many as 260 million face images. Since there is no publicly available database of similar size, it will always be difficult to compete with the FaceNet model fairly.

We directly cite the performance of the VGG Face model on YouTube Faces reported in [Parkhi et al.]. TBE-CNN* outperforms VGG Face* by as much as 2.24%. This result is consistent with results on the PaSC and COX Face databases. It is worth noting that the face representation in [Parkhi et al.] for each video frame is the average of 30 cropped patches (three scales, five random positions, and horizontal flipping). Although this strategy promotes recognition performance, it significantly reduces efficiency for real-world VFR tasks. In comparison, there are only two crops from each video frame in TBE-CNN.

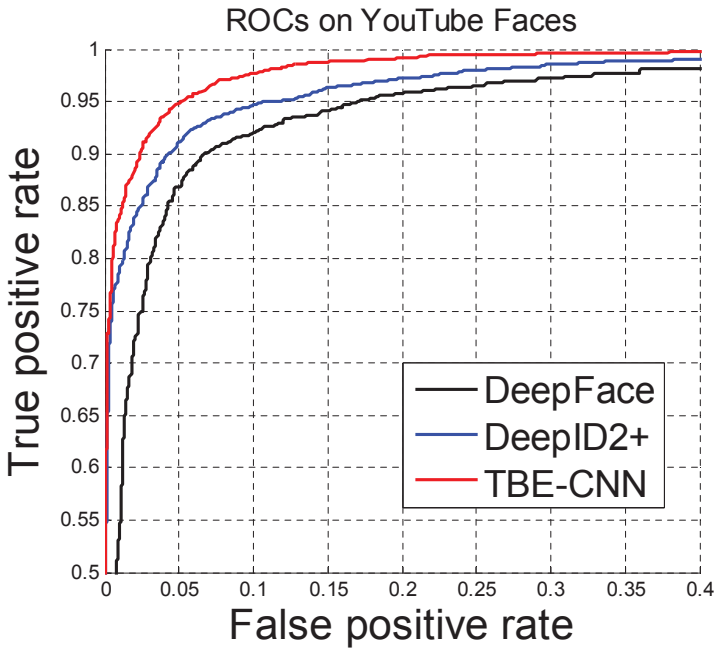


Figure 5.13: ROC curves of TBE-CNN and state-of-the-art methods on the YouTube Faces database under the "restricted" protocol.

5.7 Conclusion

VFR is a challenging task due to severe image blur, rich pose variations, and occlusion. Compared to SIFR, VFR also has more demanding efficiency requirements. Here we address these problems via a series of contributions. First, to deal with image blur, we enrich the CNN training data by applying artificial blur to make up for the shortage of real-world video training data. A training set composed of still images and their blurred versions encourages CNN to learn blur-robust representations. Second, to extract pose- and occlusion-robust representations efficiently, we propose a novel CNN architecture named TBE-CNN. TBE-CNN efficiently extracts representations of the holistic face image and facial components by sharing the low- and middle-level layers of different CNNs. It improves on single CNN model performance with only marginal increases in time and memory costs. Finally, to further promote the discriminative power of the representations learnt by TBE-CNN, we propose a novel deep metric learning approach named MDR-TL, which outperforms the widely adopted triplet loss by a considerable margin. Extensive experiments have been conducted for S2V, V2S, and

V2V tasks. Since the proposed TBE-CNN approach effectively handles image blur and pose variations, it shows clear advantages compared with state-of-the-art VFR methods on three popular video face databases.

Chapter 6

Conclusions and Future Work

There has been an explosion in face data derived from various camera devices and the internet, prompting demand for recognizing faces from image data. Face recognition has advantages in being non-invasive and easy to use. Therefore, it has become one of the most important and promising biometric techniques in many real-world systems including video surveillance, access control, forensics and security, and social networks. However, designing robust face recognition algorithms is very challenging due to: 1) small inter-personal appearance differences; and 2) large intra-personal appearance differences caused by pose, illumination, and expression variations, occlusion, and image blur.

To tackle these challenges, in this thesis we propose a series of robust face recognition algorithms that outperform existing algorithms on many popular face databases such as FERET, FRGC 2.0, CAS-PEAL-R1, LFW, CMU-PIE, Multi-PIE, PaSC, COX Face, and YouTube Faces. Since effective face representation plays a vital role in accurate face recognition, we first propose novel face image descriptors for generic face recognition purposes. Based on our research on face image descriptors, we go on to develop algorithms for the two most challenging face recognition applications: pose-invariant face recognition and surveillance video-based face recognition.

First, we propose a novel handcrafted face image descriptor named DCP. DCP efficiently encodes the second-order statistics of facial textures in the most informative directions within a face image; therefore, it has advantages in terms of both descriptive and discriminative power. We empirically find that DCP outperforms a number of existing face image descriptors with limited time and memory costs. We further extend

DCP into a comprehensive face representation scheme named MDML-DCPs. MDML-DCPs efficiently encodes invariant face image characteristics from multiple levels into patterns that are highly discriminative of inter-personal differences but robust to intra-personal variations. MDML-DCPs delivers the best performance on the challenging FERET, FRGC 2.0, CAS-PEAL-R1, and LFW databases.

We next develop a deep learning-based face image descriptor named "Multimodal Deep Face Representation" (MM-DFR). MM-DFR automatically learns face representation from multimodal image data. In MM-DFR, CNNs are designed to extract multimodal information from the original holistic face image, the frontal pose image rendered by a 3D model, and uniformly sampled image patches. The recognition ability of each CNN is promoted by carefully integrating a number of published or newly developed tricks. A feature-level fusion approach using stacked autoencoders is designed to fuse the features extracted from the set of CNNs, which is advantageous for nonlinear dimension reduction. MM-DFR achieves greater than 99% recognition accuracy on LFW using a publicly available training set.

Moreover, we extend our research on generic face recognition to PIFR. Based on our research on handcrafted face image descriptors, we propose the powerful PBPR-MtFTL framework, which is capable of handling the full range of pose variations within $\pm 90^{\circ}$ of yaw. The framework contains two parts: the first is Patch-based Partial Representation (PBPR) and the second is Multi-task Feature Transformation Learning (MtFTL). PBPR transforms the original PIFR problem into a partial frontal face recognition problem. A robust patch-based face representation scheme is developed to represent the synthesized partial frontal faces. For each patch, a transformation dictionary is learnt under the MtFTL scheme. The transformation dictionary transforms the features of different poses into a discriminative subspace in which face matching is performed. The complete PBPR-MtFTL framework outperforms state-of-the-art PIFR methods on the FERET, CMU-PIE, and Multi-PIE databases.

Finally, based on our research on deep learning-based face image descriptors, we tackle the most challenging VFR problem. We design a novel framework named TBE-CNN to handle the VFR challenges under surveillance conditions. Three major challenges are considered: image blur, occlusion, and pose variation. We make a series of contributions. First, to learn blur-robust face representations, we introduce

artificial blur to training data composed of clear still images. This strategy effectively overcomes the shortfall in real-world video training data and encourages CNN to learn blur-insensitive features automatically. Second, to enhance robustness of CNN features to pose variations and occlusion, we propose the TBE-CNN model, which extracts complementary information from the holistic face image and patches cropped around facial components. We make TBE-CNN efficient by sharing the low- and middle-level convolutional layers among multiple networks. Third, we propose a new deep metric learning model to enhance the discriminative power of TBE-CNN. With the proposed framework, state-of-the-art performance is achieved on three popular video face databases: PaSC, COX Face, and YouTube Faces.

In conclusion, the robust algorithms presented in this thesis successfully address a number of challenging practical problems in face recognition including pose variations, illumination variations, expression variations, occlusion, and image blur. These algorithms are superior to existing algorithms for generic face recognition, pose-invariant face recognition, and surveillance video-based face recognition tasks.

However, a lot of work still needs to be done to realize the full potential of face recognition in real-world applications. For example, the PBPR-MtFTL algorithm proposed in this thesis performs very well on face databases developed under laboratory conditions, but its power needs further examination on real-world databases for PIFR. In addition, CNN-based approaches need further study to solve the difficult PIFR problem. One key difficulty may be that we still lack large amounts of multi-pose training data captured under real-world circumstances. Furthermore, although TBE-CNN outperforms previous state-of-the-art VFR algorithms by a large margin, its performance at 0.1% FAR is still far from perfect. Therefore, more powerful solutions should be investigated to solve image blur and pose variation problems for VFR. Last but not least, the scale of most existing face databases is small; therefore, it would be useful to develop larger-scale face databases so that the performance of existing algorithms can be re-analyzed and their performance further improved.

We believe that, as an effective and promising biometric technique, face recognition will continue to receive intensive attention and be applied to more real-world computer vision systems. The series of contributions made in this thesis helps to expedite the design of robust face recognition algorithms for practical applications.

References

Btas 2016 video person recognition evaluation. URL http://www.pasc-eval.org/index.html. 150

Ramzi Abiantun, Utsav Prabhu, and Marios Savvides. Sparse feature extraction for pose-tolerant face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(10): 2061–2073, 2014. 91, 96

Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *Proc. Eur. Conf. Comput. Vis.*, pages 469–481. 2004. 8

Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006. 49, 50

Timo Ahonen, Esa Rahtu, Ville Ojansivu, and J Heikkila. Recognition of blurred faces using local phase quantization. In *Int. Conf. Pattern Recognit.*, pages 1–4, 2008. 9, 49, 50, 130

Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006. 19

Zahid Akhtar, Ajita Rattani, Abdenour Hadid, and Massimo Tistarelli. Face recognition under ageing effect: A comparative analysis. In *Proc. Int. Conf. Image Anal. Process.*, pages 309–318, 2013. 34

Shervin Rahimzadeh Arashloo. Efficient processing of mrfs for unconstrained-pose face recognition. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.*, pages 1–8, 2013. 111

Shervin Rahimzadeh Arashloo and Josef Kittler. Energy normalization for pose-invariant face recognition based on mrf model image matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(6):1274–1280, 2011. xiii, 17, 91, 107, 132

Shervin Rahimzadeh Arashloo and Josef Kittler. Fast pose invariant face recognition using super coupled multiresolution markov random fields on a gpu. *Pattern Recognit. Lett.*, 48:49–59, 2014. 110

Shervin Rahimzadeh Arashloo, Josef Kittler, and William J Christmas. Pose-invariant face recognition by matching on multi-resolution mrfs linked by supercoupling transform. *Comput. Vis. Image Underst.*, 115(7):1073–1083, 2011. 18, 110

Ahmed Bilal Ashraf, Simon Lucey, and Tsuhan Chen. Learning patch correspondences for improved viewpoint invariant face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–8, 2008. 20

Akshay Asthana, Tim K Marks, Michael J Jones, Kinh H Tieu, and M Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 937–944, 2011. 96, 107, 111, 114, 116, 117

Oren Barkan, Jonathan Weill, Lior Wolf, and Hagai Aronowitz. Fast high dimensional vector multiplication face recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1960–1967, 2013a. 123

Oren Barkan, Jonathan Weill, Lior Wolf, and Hagai Aronowitz. Fast high dimensional vector multiplication face recognition. In *Proc. Int. Conf. Comput. Vis.*, 2013b. 66

Jeremiah R Barr, Kevin W Bowyer, Patrick J Flynn, and Soma Biswas. Face recognition from video: A review. *Int. J. Pattern Recognit. Artif. Intell.*, 26(05), 2012. 130

Gaston Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000. 6

Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, 1997. 6

Lacey Best-Rowden, Brendan Klare, Joshua Klontz, and Anubhav K Jain. Video-to-video face matching: Establishing a baseline for unconstrained face recognition. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.*, pages 1–8, 2013. 127

J Ross Beveridge, Jonathon Phillips, David S Bolme, Bruce Draper, Geof H Givens, Yui Man Lui, Mohammad Nayeem Teli, Hao Zhang, W Todd Scruggs, Kevin W Bowyer, et al. The challenge of face recognition from digital point-and-shoot cameras. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.*, pages 1–8, 2013. 28, 129, 141, 142

J Ross Beveridge, Hao Zhang, et al. Report on the fg 2015 video person recognition evaluation. In *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognit.*, pages 1–8, 2015. 3, 4, 9, 12, 33, 127, 131, 144, 150

David Beymer and Tomaso Poggio. Face recognition from one example view. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 500–507, 1995. 20, 21

David J Beymer. Face recognition under varying pose. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 756–761, 1994. 14

Wei Bian and Dacheng Tao. Face subspace learning. In *Handbook of Face Recognition*, pages 51–77. 2011. 7

Manuele Bicego, Enrico Grosso, and Massimo Tistarelli. Person authentication from video of faces: a behavioral and physiological approach using pseudo hierarchical hidden markov models. In *Advances in Biometrics*, pages 113–120. 2006. 130

Soma Biswas, Gaurav Aggarwal, Patrick J Flynn, and Kevin W Bowyer. Pose-robust recognition of low-resolution face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):3037–3049, 2013. 127

Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, 2003. 23, 91

Volker Blanz, Patrick Grother, P Jonathon Phillips, and Thomas Vetter. Face recognition based on frontal views generated from non-frontal images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 2, pages 454–461, 2005. 91

Alfred M Bruckstein, Robert J Holt, Thomas S Huang, and Arun N Netravali. Optimum fiducials under weak perspective projection. *Int. J. Comput. Vis.*, 35(3): 223–244, 1999. 96

Roberto Brunelli and Tomaso Poggio. Face recognition: Features versus templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(10):1042–1052, 1993. xiii, 14, 17

John Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, (6):679–698, 1986. 39, 57, 97

Qiong Cao, Yiming Ying, and Peng Li. Similarity metric learning for face recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2408–2415, 2013. 123

Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2707–2714, 2010. 10

Xiujuan Chai, Shiguang Shan, Xilin Chen, and Wen Gao. Locally linear regression for pose-invariant face recognition. *IEEE Trans. Image Process.*, 16(7):1716–1725, 2007. 21, 91

Chi Ho Chan, Muhammad Atif Tahir, Josef Kittler, and Matti Pietikäinen. Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(5):1164–1177, 2013a. 63

Chi Ho Chan, Muhammad Atif Tahir, Josef Kittler, and Matti Pietikainen. Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(5):1164–1177, 2013b. 130

Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. Pcanet: A simple deep learning baseline for image classification? *arXiv preprint arXiv:1404.3606*, 2014. 10, 11

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, 2011. 59

Dong Chen, Xudong Cao, Liwei Wang, Fang Wen, and Jian Sun. Bayesian face revisited: A joint formulation. In *Proc. Eur. Conf. Comput. Vis.*, pages 566–579, 2012a. 13, 14, 62, 66, 79

Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3025–3032, 2013. xiii, 17, 41, 42, 64, 66, 119, 123

Dong Chen, Xudong Cao, David Wipf, Fang Wen, and Jian Sun. An efficient joint formulation for bayesian face verification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 131

Li-Fen Chen, Hong-Yuan Mark Liao, Ming-Tat Ko, Ja-Chen Lin, and Gwo-Jong Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern recognition*, 33(10):1713–1726, 2000. 6

Terrence Chen, Wotao Yin, Xiang Sean Zhou, Dorin Comaniciu, and Thomas S Huang. Total variation models for variable lighting face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1519–1524, 2006. 2

Yi-Chen Chen, Vishal M Patel, P Jonathon Phillips, and Rama Chellappa. Dictionary-based face recognition from video. In *Proc. Eur. Conf. Comput. Vis.*, pages 766–779, 2012b. 130

Jonghyun Choi, William Robson Schwartz, Huimin Guo, and Larry S Davis. A complementary local feature descriptor for face identification. In *Proc. IEEE Workshop Applications of Comput. Vis.*, pages 121–128, 2012. 38, 49

Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001. 20

Zhen Cui, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Image sets alignment for video-based face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2626–2633, 2012. 130

George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pages 8609–8613, 2013. 73

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proc. Int. Conf. Mach. Learn.*, pages 209–216, 2007. 13

Changxing Ding and Dacheng Tao. Robust face recognition via multimodal deep face representation. *IEEE Trans. Multimedia*, 17(11):2049–2058, 2015. 12, 25

Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Trans. Intell. Syst. Technol.*, 7:37:1–37:42, 2016. 2, 3, 4, 12, 33, 53, 85, 94, 130

Changxing Ding, Chang Xu, and Dacheng Tao. Multi-task pose-invariant face recognition. *IEEE Trans. Image Process.*, 24(3):980–993, 2015. xiii, 22, 24, 72, 73

Changxing Ding, Jonghyun Choi, Dacheng Tao, and Larry S Davis. Multi-directional multi-level dual-cross patterns for robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(3):518–531, 2016. xiii, 2, 8, 17, 70, 79, 91, 99, 128

Mika Fischer, Hazım Kemal Ekenel, and Rainer Stiefelhagen. Analysis of partial least squares for pose-invariant face recognition. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.*, pages 331–338, 2012. 24

Wen Gao, Bo Cao, Shiguang Shan, et al. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, 38(1): 149–161, 2008. 27, 34, 45

Yongsheng Gao, MKH Leung, W Wang, and Siu Cheung Hui. Fast face identification under varying pose from a single 2-d model view. *IEE Proceedings-Vision, Image and Signal Processing*, 148(4):248–253, 2001. 23

Raghuraman Gopalan, Sima Taheri, Pavan Turaga, and Rama Chellappa. A blur-robust descriptor with applications to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(6):1220–1226, 2012. 130

Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vis. Comput.*, 28(5):807–813, 2010. 27, 107

Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 498–505, 2009. 13

Abdenour Hadid and Matti Pietikäinen. Combining appearance and motion for face and gender recognition from videos. *Pattern Recognit.*, 42(11):2818–2827, 2009. 130

Hu Han, Shiguang Shan, Xilin Chen, and Wen Gao. A comparative study on illumination preprocessing in face recognition. *Pattern Recognit.*, 46(6):1691–1699, 2013. 2, 4

Mehrtash T Harandi, Conrad Sanderson, Sareh Shirazi, and Brian C Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2705–2712, 2011. 130

Tal Hassner. Viewing real-world faces in 3d. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3607–3614, 2013. 22

Huy Tho Ho and Rama Chellappa. Pose-invariant face recognition using markov random fields. *IEEE Trans. Image Process.*, 22(4):1573–1584, 2013. 91, 99, 116

Zhibin Hong, Xue Mei, Danil Prokhorov, and Dacheng Tao. Tracking via robust multi-task multi-view joint sparse representation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 649–656, 2013. 94

Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1875–1882, 2014. 131

Yiqun Hu, Ajmal S Mian, and Robyn Owens. Face recognition using sparse approximated nearest points between image sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):1992–2004, 2012. 130

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 27, 33, 34, 45, 62, 69, 79, 82, 91, 97, 107, 121, 122, 127

Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. *arXiv preprint arXiv:1512.08086*, 2015a. 132

Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Learning euclidean-to-riemannian metric for point-to-set classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1677–1684, 2014. 151, 152, 153

Zhiwu Huang, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, and Xilin Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Trans. Image Process.*, 24(12):5967–5981, 2015b. 28, 127, 129, 130, 142, 151, 152, 153

Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Face recognition on large-scale video in the wild with hybrid euclidean-and-riemannian metric learning. *Pattern Recognit.*, 48(10):3113–3124, 2015c. 150, 152, 153

Sibt Ul Hussain, Thibault Napoléon, Fréderic Jurie, et al. Face recognition using local quantized patterns. In *Proc. Int'l Conf. Biometrics*, 2012. 49, 50, 60

Wonjun Hwang, Haitao Wang, Hyunwoo Kim, Seok-Cheol Kee, and Junmo Kim. Face recognition system using multiple face model of hybrid fourier feature under uncontrolled illumination variation. *IEEE Trans. Image Process.*, 20(4):1152–1165, 2011. 63

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 147

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM Int. Conf. Multimedia*, pages 675–678, 2014. 81, 140, 143

Dalong Jiang, Yuxiao Hu, Shuicheng Yan, Lei Zhang, Hongjiang Zhang, and Wen Gao. Efficient 3d reconstruction for face recognition. *Pattern Recognit.*, 38(6):787–798, 2005. 22

Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. In *Proc. Eur. Conf. Comput. Vis.*, pages 808–821, 2012. xiii, 19

Meina Kan, Shiguang Shan, Hong Chang, and Xilin Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014. xiii, 18, 21, 116

Juho Kannala and Esa Rahtu. Bsif: Binarized statistical image features. In *Proc. IEEE Int. Conf. Pattern Recognit.*, pages 1363–1366, 2012. 10

Ira Kemelmacher-Shlizerman, Steve Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. *arXiv preprint arXiv:1512.00596*, 2015. 3, 12

Tae-Kyun Kim and Josef Kittler. Design and fusion of pose-invariant face-identification experts. *IEEE Trans. Circuits Syst. Video Technol.*, 16(9):1096–1106, 2006. 24

Tae-Kyun Kim, Josef Kittler, Hyun-Chul Kim, and Seok-Cheol Kee. Discriminant analysis by multiple locally linear transformations. In *Proc. Brit. Mach. Vis. Conf.*, volume 1, 2003. 19

Brendan F Klare, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1931–1939, 2015. 27

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1097–1105, 2012. 73

Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. Face recognition: A convolutional neural-network approach. *IEEE Trans. Neural Netw.*, 8(1):98–113, 1997. 11

Mun Wai Lee and Surendra Ranganath. Pose-invariant face recognition using a 3d deformable model. *Pattern Recognit.*, 36(8):1835–1846, 2003. 23

Zhen Lei, Matti Pietikainen, and Stan Z Li. Learning discriminant face descriptor. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(2):289–302, 2014. 10, 35, 38, 49, 50, 60, 61

Annan Li, Shiguang Shan, Xilin Chen, and Wen Gao. Maximizing intra-individual correlations for face recognition across pose differences. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 605–611, 2009. xiii, 17, 19, 94, 107

Annan Li, Shiguang Shan, and Wen Gao. Coupled bias–variance tradeoff for cross-pose face recognition. *IEEE Trans. Image Process.*, 21(1):305–315, 2012a. 107, 111, 112, 113, 115, 118, 120

Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, and Jianchao Yang. Probabilistic elastic matching for pose variant face verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3499–3506, 2013a. 122

Haoxiang Li, Gang Hua, Xiaohui Shen, Zhe Lin, and Jonathan Brandt. Eigen-pep for video face recognition. In *Proc. Asian. Conf. Comput. Vis.*, 2014a. 131

Peng Li, Yun Fu, Umar Mohammed, James H Elder, and Simon JD Prince. Probabilistic models for inference about identity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1):144–157, 2012b. 42, 43, 62, 64, 66, 123

Shaoxin Li, Xin Liu, Xiujuan Chai, Haihong Zhang, Shihong Lao, and Shiguang Shan. Morphable displacement field based image matching for face recognition across pose. In *Proc. Eur. Conf. Comput. Vis.*, pages 102–115, 2012c. 20, 107, 109, 110, 111, 116

Shaoxin Li, Xin Liu, Xiujuan Chai, Haihong Zhang, Shihong Lao, and Shiguang Shan. Maximal likelihood correspondence estimation for face recognition across pose. *IEEE Trans. Image Process.*, 23(10):4587–4600, 2014b. 109, 116

Yan Li, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Fusing magnitude and phase features for robust face recognition. In *Proc. Asian. Conf. Comput. Vis.*, pages 601–612, 2013b. 62

Shengcai Liao, Anil K Jain, and Stan Z Li. Partial face recognition: Alignment-free approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(5):1193–1205, 2013. 92, 94

Jinguo Liu, Yafeng Deng, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015. 12, 127, 132

Luoqi Liu, Li Zhang, Hairong Liu, and Shuicheng Yan. Toward large-population face identification in unconstrained videos. *IEEE Trans. Circuits Syst. Video Technol.*, 24 (11):1874–1884, 2014. 130

Xiaoming Liu and Tsuhan Chen. Pose-robust face recognition using geometry assisted probabilistic modeling. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 1, pages 502–509, 2005. 23

Zhiming Liu and Chengjun Liu. Robust face recognition using color information. In *Proc. Int'l Conf. Biometrics*, pages 122–131, 2009. 62, 63

Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou. Multi-manifold deep metric learning for image set classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1137–1145, 2015. 131

Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Trans. Image Process.*, 23 (8):3656–3670, 2014. 100, 101

Behrooz Mahasseni and Sinisa Todorovic. Latent multitask learning for view-invariant action recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3128–3135, 2013. 94

David Masip, Ágata Lapedriza, and Jordi Vitrià. Multitask learning applied to face recognition. In *1st Spanish Workshop on Biometrics*, pages 1–8, 2007. 94

David Masip, Ágata Lapedriza, and Jordi Vitrià. Multitask learning-an application to incremental face recognition. In *Proc. Int. Conf. Comput. Vis. App.*, pages 585–590, 2008. 94

Sandra Mau, Shaokang Chen, Conrad Sanderson, and Brian C Lovell. Video face matching using subset selection and clustering of probabilistic multi-region histograms. *arXiv preprint arXiv:1303.6361*, 2013. 130

Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2666–2672, 2012. 13

Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(12):2262–2275, 2010. 97

Masashi Nishiyama, Abdenour Hadid, Hidenori Takeshima, Jamie Shotton, Tatsuo Kozakaya, and Osamu Yamaguchi. Facial deblur inference using subspace analysis for recognition of blurred faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(4): 838–845, 2011. 4, 130, 134

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002. 49, 50

Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, and Xueming Li. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. *arXiv preprint arXiv:1409.5114*, 2014. 16

Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 11, 127, 131, 149, 150, 151, 152, 153, 154

Omkar M Parkhi, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. A compact and discriminative face track descriptor. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1693–1700, 2014. 131

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proc. IEEE*

Int. Conf. Advanced Video and Signal Based Surveillance, pages 296–301, 2009. 76, 108

Alex Pentland, Baback Moghaddam, and Thad Starner. View-based and modular eigenspaces for face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 84–91, 1994. xiii, 14, 17

Jonathon Phillips, J Ross Beveridge, David S Bolme, Bruce Draper, Geof H Givens, Yui Man Lui, Su Cheng, Mohammad Nayeem Teli, Hao Zhang, et al. On the existence of face quality measures. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.*, pages 1–8, 2013. 130

P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, 2000. 26, 34, 44, 47, 107

P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *Proc. Comput. Vis. Pattern Recognit.*, volume 1, pages 947–954, 2005. 26, 34, 45

Simon JD Prince and James H Elder. Creating invariance to" nuisance parameters" in face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 2, pages 446–453, 2005. 19

Simon JD Prince and James H Elder. Probabilistic linear discriminant analysis for inferences about identity. In *Proc. Int. Conf. Comput. Vis.*, pages 1–8, 2007. 13, 42, 79, 122

Simon JD Prince, J Warrell, James H Elder, and Fatima M Felisberti. Tied factor analysis for face recognition across large pose differences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):970–984, 2008. xiii, 19, 24, 42, 91

Aruni RoyChowdhury, Tsung-Yu Lin, Subhransu Maji, and Erik Learned-Miller. Face identification with bilinear cnns. *arXiv preprint arXiv:1506.01342*, 2015. 131

Mark Schmidt. The minfunc package. URL `http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`. 103

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 815–823, 2015. 11, 12, 24, 80, 81, 87, 127, 129, 131, 137, 138, 145, 146, 154

Mang Shao, Danhang Tang, Yang Liu, and Tae-Kyun Kim. A comparative study of video-based object recognition from an egocentric viewpoint. *Neurocomputing*, 2015. 130

Abhishek Sharma and David W Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 593–600, 2011. 19

Abhishek Sharma, Abhishek Kumar, H Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2160–2167, 2012. xiii, 19, 94

Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12):1615–1618, 2003. 3, 107

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 11, 70, 72, 73

Karen Simonyan, Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Fisher vector faces in the wild. In *Proc. Brit. Mach. Vis. Conf.*, pages 1–12, 2013. 66, 122, 123

Yu Su, Shiguang Shan, Xilin Chen, and Wen Gao. Hierarchical ensemble of global and local classifiers for face recognition. *IEEE Trans. Image Process.*, 18(8):1885–1896, 2009. 63

Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3476–3483, 2013a. 108, 142

Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1988–1996, 2014. 7, 11, 12, 70, 71, 73, 74, 84, 86, 87, 128, 131, 132, 137

Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2892–2900, 2015. 73, 87, 127, 154

Zhan-Li Sun, Kin-Man Lam, and Qing-Wei Gao. Depth estimation of face images using the nonlinear least-squares model. *IEEE Trans. Image Process.*, 22(1):17–30, 2013b. 96

Jinli Suo, Song-Chun Zhu, Shiguang Shan, and Xilin Chen. A compositional and dynamic model for face aging. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):385–401, 2010. 3

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–9, 2015. 11, 135

Yaniv Taigman, Lior Wolf, Tal Hassner, et al. Multiple one-shots for utilizing class label information. In *Proc. Brit. Mach. Vis. Conf.*, pages 1–12, 2009. 123

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1701–1708, 2014a. 11, 41, 70, 71, 73, 82, 84, 87, 131, 154

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. *arXiv preprint arXiv:1406.5266*, 2014b. 84

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2746–2754, 2015. 24

Xiaoyang Tan and Bill Triggs. Fusing gabor and lbp feature sets for kernel-based face recognition. In *Proc. Int'l Conf. Analysis and Modeling of Faces and Gestures*, pages 235–249, 2007. 60, 63

Xiaoyang Tan and Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.*, 19(6):1635–1650, 2010. 2, 8, 45, 49, 50, 63, 91, 108

Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Fuyan Zhang. Face recognition from a single image per person: A survey. *Pattern Recognit.*, 39(9):1725–1745, 2006. 4, 13

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 19

Jirí Trefnỳ and Jirí Matas. Extended set of local binary patterns for rapid object detection. In *Proc. Comput. Vis. Winter Workshop*, 2010. 8, 49, 50

Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neurosci.*, 3(1):71–86, 1991. 6

Sibt Ul Hussain and Bill Triggs. Visual recognition using local quantized patterns. In *Proc. Eur. Conf. Comput. Vis.*, pages 716–729, 2012. 10, 33

Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence*. Springer, 1996. 105

Paul Viola and Michael J Jones. Robust real-time face detection. *Int. J. Comput. Vis.*, 57(2):137–154, 2004. 27

Ngoc-Son Vu and Alice Caplier. Enhanced patterns of oriented edge magnitudes for face recognition and image matching. *IEEE Trans. Image Process.*, 21(3):1352–1365, 2012. 8, 49, 50

Dayong Wang, Charles Otto, and Anil K Jain. Face search at scale: 80 million gallery. *arXiv preprint arXiv:1507.07242*, 2015. 83, 84, 87

Haitao Wang, Stan Z Li, and Yangsheng Wang. Face recognition under varying lighting conditions using self quotient image. In *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognit.*, pages 819–824, 2004. 2

Ruxin Wang and Dacheng Tao. Recent progress in image deblurring. *arXiv preprint arXiv:1409.6838*, 2014. 133, 134

Xiaogang Wang and Xiaoou Tang. A unified framework for subspace face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1222–1228, 2004. 6

Xiaogang Wang and Xiaoou Tang. Random sampling for subspace face recognition. *Int. J. Comput. Vis.*, 70(1):91–104, 2006. 6

Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1473–1480, 2005. 12

Laurenz Wiskott, J-M Fellous, N Kuiger, and Christoph Von Der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):775–779, 1997. xiii, 17

Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 529–534, 2011a. 27, 129, 142, 154

Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(10):1978–1990, 2011b. 45, 62, 130

John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009. 3, 13

Shufu Xie, Shiguang Shan, Xilin Chen, Xin Meng, and Wen Gao. Learned local gabor patterns for face representation and recognition. *Signal Process.*, 89(12):2333–2344, 2009. 61

Shufu Xie, Shiguang Shan, Xilin Chen, and Jie Chen. Fusing local patterns of gabor magnitude and phase for face recognition. *IEEE Trans. Image Process.*, 19(5):1349–1361, 2010. 2, 7, 9, 34, 39, 49, 50, 60, 63

Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proc. Comput. Vis. Pattern Recognit.*, pages 532–539, 2013. 41

Huan Xu and Shie Mannor. Robustness and generalization. *Mach. learn.*, 86(3):391–423, 2012. 104

Jian Yang, David Zhang, Alejandro F Frangi, and Jing-yu Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):131–137, 2004. 6

Dong Yi, Zhen Lei, and Stan Z Li. Towards pose robust face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3539–3545, 2013. xiii, 17, 75

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. xii, 2, 69, 70, 71, 72, 73, 79, 80, 82, 83, 84, 143

Baochang Zhang, Shiguang Shan, Xilin Chen, and Wen Gao. Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. *IEEE Trans. Image Process.*, 16(1):57–68, 2007a. 61

Lun Zhang, Rufeng Chu, Shiming Xiang, Shengcai Liao, and Stan Z Li. Face detection based on multi-block lbp representation. In *Advances in biometrics*, pages 11–18. 2007b. 9

Tianhao Zhang, Dacheng Tao, Xuelong Li, and Jie Yang. Patch alignment for dimensionality reduction. *IEEE Trans. Knowl. Data Eng.*, 21(9):1299–1313, 2009. 112, 145, 146

Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Proc. Int. Conf. Comput. Vis.*, volume 1, pages 786–791, 2005. 9, 61

Xiaozheng Zhang, Yongsheng Gao, and Maylor KH Leung. Automatic texture synthesis for face recognition from single views. In *Proc. IEEE Int. Conf. Pattern Recognit.*, volume 3, pages 1151–1154, 2006. 23

Yizhe Zhang, Ming Shao, Edward K Wong, and Yun Fu. Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2416–2423, 2013. xiii, 18, 21

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Proc. Eur. Conf. Comput. Vis.*, pages 94–108, 2014. 72

Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):915–928, 2007. 9

Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003. 4

Erjin Zhou, Zhimin Cao, and Qi Yin. Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015. xv, 70, 80, 81

Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Simon Chi-Keung Shiu, and Dejing Zhang. Image set-based collaborative representation for face recognition. *IEEE Trans. Inf. Forensics Security*, 9(7):1120–1132, 2014a. 130

Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity-preserving face space. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 113–120, 2013. xiii, 18, 21, 91, 108, 113, 114, 115, 116, 119

Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: A deep model for learning face identity and view representations. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 217–225, 2014b. 94, 107, 116, 117, 119