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 ±90° 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

Contents viii

List of Figures xii

List of Tables xx

1 Introduction 1

1.1 Background . . . 1

1.2 Facial Feature Extraction . . . 5

1.2.1 Subspace Learning-based Representations . . . 5

1.2.2 Local Descriptor-based Representations . . . 7

1.2.3 Deep Learning-based Representations . . . 11

1.3 Classification Models . . . 12

1.3.1 Discriminative Models . . . 12

1.3.2 Generative Models . . . 13

1.4 Pose-invariant Face Recognition . . . 14

1.4.1 Pose-robust Feature Extraction . . . 17

1.4.2 Multi-view Subspace Learning . . . 19

1.4.3 Face Synthesis Based on 2D Methods . . . 20

1.4.4 Face Synthesis Based on 3D Methods . . . 22

1.4.5 Hybrid Methods . . . 23

1.4.6 Relationships Between the Four Categories . . . 24

1.5 Face Databases . . . 25

1.5.1 Still Face Image Databases . . . 26

1.5.2 Video Face Databases . . . 27

1.6 Contributions and Related Publications 28

2 Multi-Directional Multi-Level Dual-Cross Patterns 32

2.1 Introduction 33

2.2 Dual-Cross Patterns 34

2.2.1 Local Sampling 35

2.2.2 Pattern Encoding 36

2.2.3 Dual-Cross Grouping 37

2.2.4 DCP Face Image Descriptor 37

2.3 Multi-Directional Multi-Level Dual-Cross Patterns 39

2.3.1 The MDML-DCPs Scheme 39

2.3.2 Implementation Details 41

2.4 Face Recognition Algorithm 42

2.4.1 WPCA 42

2.4.2 PCA combined with PLDA 43

2.5 Experiments 44

2.5.1 Empirical Justification for Dual-Cross Grouping 46

2.5.2 Parameter Selection of DCP 47

2.5.3 Evaluation of the Performance of DCP 49

2.5.4 The Contribution of Multi-directional Filtering 57

2.5.5 Performance Evaluation of MDML-DCPs 58

2.6 Conclusion 64

3 Multimodal Deep Face Representation 67

3.1 Introduction 67

3.2 Related Studies 70

3.2.1 Face Image Representation 70

3.2.2 Multimodal-based Face Recognition 70

3.3 Multimodal Deep Face Representation 72

3.3.1 Single CNN Architecture 72

3.3.2 Combination of CNNs using Stacked Auto-Encoder 76

3.4 Face Matching with MM-DFR 78

3.5 Experiments 79

3.5.1 Performance Comparison with Single CNN Model 82
3.5.2 Performance of the Eight CNNs in MM-DFR 83
3.5.3 Fusion of CNNs with SAE 85
3.5.4 Performance of MM-DFR with Joint Bayesian 86
3.5.5 Face Identification on CASIA-WebFace Database 87
3.6 Conclusion 88

4 Multi-task Pose-Invariant Face Recognition 90
4.1 Introduction 91
4.2 Related Studies 94
4.3 Face Representation for the Pose Problem 94
4.3.1 Face Pose Normalization 96
4.3.2 Unoccluded Facial Texture Detection 96
4.3.3 Patch-based Face Representation 97
4.4 Multi-task Feature Transformation Learning 99
4.4.1 Feature Transformation Learning 100
4.4.2 Iterative Optimization Algorithm 102
4.4.3 Theoretical Analysis 104
4.5 Face Matching with PBPR-MtFTL 106
4.6 Experimental Evaluation 107
4.6.1 Comparison on CMU-PIE and FERET 109
4.6.2 Comparison with Single-task Baselines 109
4.6.3 Recognition across Pose and Illumination 113
4.6.4 Recognition across Pose and Recording Session 114
4.6.5 Recognition across Pose, Illumination, and Recording Session 117
4.6.6 Parameter Evaluation for MtFTL 118
4.6.7 Performance in the Fully-Automatic Mode 120
4.6.8 Extension to Unconstrained Face Verification 121
4.7 Conclusion 123
4.8 Proof of Theorem 1 124
4.9 Proof of Theorem 2 124

5 Trunk-Branch Ensemble Convolutional Neural Networks for Video-based Face Recognition 126

5.1 Introduction . . . 127

5.2 Related Works . . . 129

5.2.1 Video-based Face Recognition . . . 129

5.2.2 Deep Learning Methods for Face Recognition . . . 131

5.3 Trunk-Branch Ensemble CNN for VFR . . . 132

5.3.1 Artificially Simulated Video Data . . . 133

5.3.2 Trunk-Branch Ensemble CNN . . . 134

5.4 TBE-CNN Training . . . 137

5.4.1 Mean Distance Regularized Triplet Loss . . . 137

5.5 VFR with TBE-CNN . . . 140

5.6 Experiments . . . 141

5.6.1 Implementation Details of TBE-CNN . . . 143

5.6.2 Effectiveness of Simulated Video Training Data . . . 143

5.6.3 Effectiveness of MDR-TL . . . 145

5.6.4 Effectiveness of Trunk-Branch Fusion . . . 147

5.6.5 Performance Comparison on PaSC . . . 149

5.6.6 Performance Comparison on COX Face . . . 153

5.6.7 Performance Comparison on YouTube Faces . . . 153

5.7 Conclusion . . . 155

6 Conclusions and Future Work 157

References 160

List of Figures

1.1 A typical face recognition system pipeline. The image is selected from the CASIA-WebFace database [Yi et al., 2014]. 2

1.2 The two typical factors that affect facial appearance: variations exhibited by the face itself and variations caused by imaging conditions. . 4

1.3 The evolution of facial feature extraction methods. 5

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. 7

1.5 The principle of LBP-based face representation. 8

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. 10

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. 15

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. 16

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]. 17

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]. 18

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. 19

1.12 Three main 2D-based pose normalization schemes. (a) Piece-wise warping; (b) patch-wise warping; and (c) pixel-wise displacement. . . 21

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. 22

1.14 The evolution of face databases. 26

1.15 Structure of this thesis. 29

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}. 35

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. 38

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. . 40

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. 41

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. . 44

2.6 Sample images from LFW. Images in the two rows are aligned by a similarity transformation and an affine transformation, respectively. . 44

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. 46

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. 47

2.9 The mean rank-1 identification rates of DCP and LBP on four FERET probe sets as a function of N. 48

2.10 Performance comparison between DCP and MsLBP on the four face datasets. 57

2.11 ROC curves of the MDML-DCPs method and other state-of-the-art methods in the unrestricted paradigm. 65

3.1 Face images in real-world applications usually exhibit rich variations in pose, illumination, expression, and occlusion. 68

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. 69

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. 76

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. . 77

3.5 More examples about the uniformly detected landmarks that are projected from a generic 3D face model to 2D images. 77

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. . . . 81

3.7 Performance comparison on LFW with different usage strategies of ReLU nonlinearity. 83

3.8 ROC curves of different usage strategies of the ReLU nonlinearity on LFW. 85

3.9 Performance comparison between the proposed MM-DFR approach and single modality-based CNN on the face verification task. 87

3.10 CMS curves by different combinations of modalities on the face identification task. 89

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. 92

4.2 Overview of the proposed PBPR-MtFTL framework for pose-invariant face recognition, as applied to the recognition of arbitrary pose probe faces. 93

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. 95

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. 98

4.5 Examples of facial contour detection for unconstrained face images in the LFW dataset. 98

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. 100

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 = $\pm 60°$. 112

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 $+90°$. 113

4.9 Performance comparison of different methods on combined variations of pose and recording session. 117

4.10 Performance comparison of different methods on combined variations of pose, illumination, and recording session. 118

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. 120

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. 121

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. 122

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. 128

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). 133

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. 135

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. 138

5.5 Illustration of TBE-CNN training with MDR-TL. MDR-TL is employed to further enhance the discriminative power of learnt face representations. 138

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. 141

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. 145

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. 146

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. 147

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. 148

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. 148

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. 149

5.13 ROC curves of TBE-CNN and state-of-the-art methods on the YouTube Faces database under the “restricted” protocol. 155

List of Tables

2.1 Feature Size of the Investigated Face Image Descriptors 50

2.2 Identification Rates for Different Descriptors on FERET 52

2.3 Rank-1 Identification Rates for Different Face Image Descriptors on the Nine Probe Sets of PEAL 54

2.4 Verification Results on the FRGC 2.0 Experiment 1 55

2.5 Verification Results on the FRGC 2.0 Experiment 4 55

2.6 Mean Verification Accuracy on the LFW View 2 Data 56

2.7 Identification Rates for Different Methods on FERET 60

2.8 Rank-1 Identification Rates for Different Methods on the Nine Probe Sets of PEAL 61

2.9 Verification Rates at 0.1% FAR for Different Methods on the FRGC 2.0 Experiments 1 and 4 63

2.10 Mean Verification Accuracy on the LFW View 2 Data 66

3.1 Details of the model architecture for NN1 74

3.2 Details of the model architecture for NN2 75

3.3 Performance Comparison on LFW using Single CNN Model on Holistic Face Image 84

3.4 Performance Comparison on LFW of Eight Individual CNNs 84

3.5 Performance Evaluation of MM-DFR with JB 87

3.6 The rank-1 identification rates by Different Combinations of Modalities on CASIA-WebFace Database 88

4.1 Model Parameters Estimated on the Validation Subsets for Different Databases 109

4.2 Performance Comparison with State-of-the-art PIFR Methods on CMU-PIE . . . 110
4.3 Performance Comparison with State-of-the-art PIFR Methods on FERET 111
4.4 Rank-1 Identification Rates on Combined Variations of Pose and Illumination on Multi-PIE . . . 115
4.5 Rank-1 Identification Rates on Combined Variations of Pose and Recording Session on Multi-PIE . . . 116
4.6 Rank-1 Identification Rates on Combined Variations of Pose, Illumination, and Recording Session on Multi-PIE . . . 119
4.7 Performance comparison on LFW with state-of-the-art methods based on single face representation . . . 123

5.1 Trunk Network Parameters (GoogLeNet) . . . 136
5.2 Verification Rates (%) at 1% FAR on PaSC with Different Types of Training Data . . . 145
5.3 Verification Rates (%) at 1% FAR of Different Methods on PaSC . . . 150
5.4 Rank-1 Identification Rates (%) under the V2S/S2V Settings for Different Methods on the COX Face Database . . . 151
5.5 Rank-1 Identification Rates (%) under the V2V Setting for Different Methods on the COX Face Database . . . 152
5.6 Mean Verification Accuracy on the YouTube Faces Database (Restricted Protocol) . . . 154