INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY DELHI

Unraveling Representations for Face Recognition: from Handcrafted to Deep Learning

by

Gaurav Goswami

Under the supervision of
Dr. Richa Singh
Dr. Mayank Vatsa

Indraprastha Institute of Information Technology Delhi
November, 2018

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY DELHI

Unraveling Representations for Face Recognition: from Handcrafted to Deep Learning

by

Gaurav Goswami

Submitted
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

to the
Indraprastha Institute of Information Technology Delhi
November, 2018

Certificate

This is to certify that the thesis titled "**Unraveling Representations for Face Recognition: from Handcrafted to Deep Learning**" being submitted by **Gaurav Goswami** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

November, 2018
Dr. Richa Singh
Associate Professor

November, 2018
Dr. Mayank Vatsa
Associate Professor

Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgment

First and foremost, I would like to acknowledge the tremendous support and understanding of my advisors, Dr. Richa Singh and Dr. Mayank Vatsa, without whose guidance this would not have been possible. They have been better advisors than I have ever hoped for and they have made this journey instructive, memorable, discriminative, and feature-rich. I would like to thank Dr. Ratha, who hosted me at the TJ Watson Research Center and provided me an opportunity to work with him on multiple papers and intellectual property disclosures. He continues to be a positive influence in my academic and professional career. I would like to thank Dr. Noore for hosting me at the West Virginia University and helping me get into the right state of mind at the onset of my academic career. I would also like to thank Dr. Angshul Majmudar for his knowledge-dense guidance during my research in sparse learning. I would like to acknowledge my seniors in the IAB lab who have been gracious in sharing their knowledge and experience with me and helped and guided me at every step of my academic and professional journey: Dr. Samarth, Dr. Tejas, Dr. Anush, Dr. Himanshu, and Dr. Brian. I must also express my gratitude to all of my co-authors who have made conducting research with them fun and exciting: Paritosh, Akshay, Romil, Anurag, Aishwarya, Pawas, and Ekam. I cannot dare not mention the rest of IAB lab which has been a second home and a support structure through thick and thin over the years: Soumyadeep, Maneet, Shruti, Rohit, Aakarsh, Saheb, Naman, and Daksha.

I would like to thank Dr. Pankaj Jalote for giving me the opportunity to be the first student to roll-over from the B.Tech program to the Ph.D. program which enabled me to continue my academic relationship with my advisors. I would like to thank Priti ma'am and Sheetu ma'am for accommodating and helping with all of my administrative requests. I would like to thank IIIT-Delhi, IBM, and the Ministry of Electronics and Information Technology for their financial support during my Ph.D and funding my research. I would also like to thank the encouragement and support of my parents who stood with me every step of the way with confidence and patience. I would like to thank all the reviewers who have provided me valuable feedback on my many submissions and helped shape each chapter of this dissertation to be better and more comprehensive. Finally, I would like to thank and acknowledge the support and encouragement of my lovely wife who spent the time and effort to patiently understand my line of work and academic lifestyle. To conclude this section, I would like to dedicate this dissertation to all of the people mentioned above since each and everyone of them was an integral and crucial part of this journey.

Unraveling Representations for Face Recognition: from Handcrafted to Deep Learning

by

Gaurav Goswami

Abstract

Automatic face recognition in unconstrained environments is a popular and challenging research problem. With the improvements in recognition algorithms, focus has shifted from addressing various covariates individually to performing face recognition in truly unconstrained scenarios. Face databases such as the YouTube Faces and the Point-and-shoot-challenge capture a wide array of challenges such as pose, expression, illumination, resolution, and occlusion simultaneously. In general, every face recognition algorithm relies on some form of feature extraction mechanism to succinctly represent the most important characteristics of face images so that machine learning techniques can successfully distinguish face images of one individual apart from those of others. This dissertation proposes novel feature extraction and fusion paradigms along with improvements to existing methodologies in order to address the challenge of unconstrained face recognition. In addition, it also presents a novel methodology to improve the robustness of such algorithms in a generalizable manner.

We begin with addressing the challenge of utilizing face data captured from consumer level RGB-D devices to improve face recognition performance without increasing the operational cost. The images captured using such devices is of poor quality compared to specialized 3D sensors. To solve this, we propose a novel feature descriptor based on the entropy of RGB-D faces along with the saliency feature obtained from a 2D face. Geometric facial attributes are also extracted from the depth image and face recognition is performed by fusing both the descriptor and attribute match scores. While score level fusion does increase the robustness of the overall framework, it cannot take into account and utilize the additional information present at the feature level. To address this challenge, we need a better feature-level fusion algorithm that can combine multiple features while preserving as much of this information before the score computation stage. To accomplish this, we propose the Group Sparse Representation based Classifier (GSRC) which removes the requirement for a separate feature-level fusion mechanism and integrates multiple features seamlessly into classification. We also propose a kernelization based extension to the GSRC that further improves its ability to separate classes that have high inter-class similarity.

We next address the problem of efficiently using large amount of video data to perform face recognition. A single video contains hundreds of images, however, not all frames of a video contain useful features for face recognition and some frames might even deteriorate performance. Keeping this in mind, we propose a novel face verification algorithm which starts with selecting feature-rich frames from a video sequence using discrete wavelet transform and entropy computation. Frame selection is followed by learning a joint representation from the proposed deep learning architecture which is a combination of stacked denoising sparse autoencoder and deep Boltzmann machine. A multilayer neural network is used as classifier to obtain the verification decision.

Currently, most of the highly accurate face recognition algorithms are based on deep learning based feature extraction. These networks have been shown in literature to be vulnerable to engineered adversarial attacks. We assess that non-learning based image-level distortions can also adversely affect the performance of such algorithms. We capitalize on how some of these errors propagate through the network to devise detection and mitigation methodologies that can help improve the real-world robustness of deep network based face recognition. The proposed algorithm does not require any re-training of the existing networks and is not specific to a particular type of network. We also evaluate the generalizability and efficacy of the approach by testing it with multiple networks and distortions. We observe favorable results that are consistently better than existing methodologies in all the test cases.

Table of Contents

1 Introduction **1**

1.1 Face Recognition: Progress from 2007 to 2018 5

1.1.1 2007-2013: Fusion of Handcrafted Features and Distance Metric Learning 7

1.1.2 2014-2018: A Focus on Deep Learning 12

1.2 Research Contributions 17

2 RGB-D Face Recognition with Texture and Attribute Features **21**

2.1 Introduction 21

2.2 Proposed RGB-D Face Recognition Algorithm 24

2.2.1 Preprocessing 24

2.2.2 RISE: RGB-D Image descriptor based on Saliency and Entropy 24

2.2.3 ADM: Attributes based on Depth Map 31

2.2.4 Combining RISE and ADM 33

2.3 Experimental Results 34

2.3.1 Database and Experimental Protocol 34

2.3.2 Results and Analysis: Experiment 1 37

2.3.3 Results and Analysis: Experiment 2 43

2.3.4 Experiments on 3D TEC dataset 46

2.4 Summary 47

3 Group Sparse Representation Based Classification for Multi-feature Multimodal Biometrics **49**

3.1 Preliminaries 53

3.1.1 Sparse Representation based Classification 53
3.1.2 Block/Joint Sparse Classification 55
3.1.3 Non-Linear Extensions 57
3.1.4 Extension to Multiple Measurement Vectors 58
3.2 Proposed Group Sparse Representation Based Classification 59
3.2.1 Multi-feature Classification 62
3.2.2 Combined Multimodal and Multi-feature Classification 63
3.3 Proposed KGSRC Algorithm 65
3.4 Experimental Results and Analysis 68
3.4.1 Algorithms used for Performance Evaluation 68
3.4.2 Results and Analysis 69
3.5 KGSRC: Case Studies 77
3.5.1 Cross-distance Face Recognition 79
3.5.2 RGB-D Face Recognition 79
3.5.3 Multi-modal Biometric Recognition 80
3.6 KGSRC: Experimental Results and Analysis 80
3.6.1 Cross-distance Face Recognition in Surveillance 80
3.6.2 RGB-D Face Recognition 84
3.6.3 Multi-modal Biometric Recognition 85
3.7 Summary 87

4 Face Verification via Learned Representation on Feature-Rich Video Frames 89
4.1 Proposed Face Recognition Algorithm 95
4.1.1 Entropy based Frame Selection 95
4.1.2 Deep Learning Framework for Feature Extraction 98
4.1.3 Face Verification using Feature Richness and Deep Learning based Representation 105
4.2 Results and Analysis 107
4.2.1 Database and Experimental Protocol 107
4.2.2 Experimental Results 108

4.2.3 Comparison with Recent CNN based Algorithms 117

4.3 Summary 118

5 Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks 119

5.1 Adversarial Attacks on Deep Learning based Face Recognition 121

5.1.1 Image-level Distortions 122

5.1.2 Face-level Distortions 123

5.1.3 Learning based Adversaries 124

5.2 Adversarial Distortions: Results and Analysis 124

5.2.1 Databases 125

5.2.2 Existing Networks and Systems 125

5.2.3 Results and Analysis 126

5.3 Detection and Mitigation of Adversarial Attacks 127

5.3.1 Network Analysis and Detection 128

5.3.2 Mitigation: Selective Dropout 131

5.3.3 Experimental Protocol 132

5.3.4 Results and Analysis 133

5.4 Summary 136

6 Conclusions and Future Work 137

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

1-1 Illustrating different covariates that deter the accuracy of face recognition algorithms. Images are taken from the AR face database [141], the CMU Multi-PIE database [69], the SCFace database [68], the Large Age Gap (LAG) database [18], and the KaspAROV database [30]. 2

1-2 Illustrating the scope of this dissertation. Using multiple modes of input such as single face image, depth image in addition to face image, and face video, we solve the challenges in extracting efficient features pertaining to the covariates of pose, illumination, expression, cross-view, and occlusion. We also propose adversarial detection and mitigation techniques to ensure the robustness of the representations in a real world scenario. 6

1-3 Illustrating how a deep network extracts facial features. Images (a) to (d) represent the output of progressively deeper convolution layers from the VGG deep network. Images (e) to (h) represent outputs from convolution layers of equivalently increasing depth from the LightCNN network. As we can see, the first layer of output consists of basic patterns primarily simulating simple edge detection that continue to evolve into more complex combinations of edges and texture in the later layers of the network. The learned features are optimized to obtain the best recognition performance and therefore capture information that focuses on preserving the differentiability of different faces. 11

1-4 Features can be extracted from various modalities of face images. While a traditional RGB (2D image) provides only color and texture information, it may be combined with depth information to obtain a RGB-D based representation, or may be combined with a time dimension (by capturing the same object across multiple frames separated by acquisition time) to obtain a video based representation. 18

2-1 Different modes of capture: (a) RGB image, (b) depth map captured using Kinect, and (c) Range image from 3D TEC dataset [206] obtained using a 3D scanner. . . . 22

2-2 Illustrating the steps involved in the proposed RGB-D face recognition algorithm. . 22

2-3 RGB-D images of two subjects illustrating the inter-class similarities of RGB images and depth maps. 26

2-4 Illustrating the steps of the proposed RISE algorithm. 26

2-5 Steps involved in the proposed ADM approach. 29

2-6 Sample images of a subject in two sessions from the IIIT-D RGB-D database. . . . 35

2-7 Analyzing the proposed RISE algorithm and its individual components on the IIIT-D RGB-D face database. 39

2-8 Analyzing the performance of different combinations of the proposed algorithm with 3D PCA and fusion algorithms on the IIIT-D RGB-D face database. 40

2-9 Analyzing the effect of weights in match score level fusion using weighted sum rule on the IIIT-D RGB-D face database. 40

2-10 Analyzing the effect of weights in rank level fusion using weighted borda count on the IIIT-D RGB-D face database. 40

2-11 Analyzing the effect of gallery size on the identification performance on the IIIT-D RGB-D face database. 41

2-12 Comparing the performance of the proposed approach with existing 2D and 3D approaches on the extended database. 41

2-13 Analyzing the performance of the proposed algorithm. The first row (Case 1) presents sample gallery and probe images when all the algorithms are able to recognize. The second row (Case 2) presents example gallery and probe images when only the proposed algorithm is able to correctly identify the subject at rank-1. . . . 46

2-14 Comparing the identification performance of the proposed algorithm with COTS on all three databases. 47

3-1 A concept diagram of the proposed algorithm. 52

3-2 Illustrating the Sparse Inverse Problem 55

3-3 Illustrating the proposed GSRC algorithm. 60

3-4 CMC curves on the WVU multimodal database: individual features, SRC and GSRC on (a) face, (b) fingerprint and (c) iris. 70

3-5 CMC curves on the LEA multimodal database: individual features, SRC and GSRC on (a) face, (b) fingerprint,and (c) iris. 71

3-6 CMC curves for identification on the WVU multimodal database. Comparing the performance of (a) traditional SRC with the proposed GSRC when two modalities are considered at a time and (b) GSRC, SRC, sum rule, and context switching algorithms when all three modalities are considered at a time. 74

3-7 CMC curves for identification on the LEA multimodal database. Comparing the performance of (a) traditional SRC with the proposed GSRC when two modalities are considered at a time and (b) GSRC, SRC, sum rule, and context switching algorithms when all three modalities are considered at a time. 75

3-8 Flow diagrams for each of the three biometric problems that we use to evaluate the proposed KGSRC algorithm. (a) Cross-distance face recognition (b) RGB-D face recognition (c) Multi-modal biometric recognition 78

3-9 Identification results on the SCFace database at three different distances of the probe image: (a) Distance 1 = 4.2 metres, (b) Distance 2 = 2.6 metres, and (c) Distance 3 = 1.0 metres. The matching is performed between a high distance probe and a low distance (high resolution) gallery. For both the KGSRC and GSRC algorithms, the best feature combination is showcased in the CMC curve. 81

3-10 Identification results on the two RGB-D databases when compared to GSRC. For both the KGSRC and GSRC algorithms, the best feature combination is showcased in the CMC. We see that kernelization using the proposed algorithm improves the performance. 82

4-1 A subset of frames illustrating the amount of information present in a video. A single video can capture a subject's face under different pose, expression, and illumination variations. While some frames can be highly useful for face recognition, others can be detrimental to performance. Images are frames from the PaSC database [12]. 90

4-2 Summarizing the performance of some of the best performing face verification algorithms on the YTF [114]. It is evident that there is a huge gap in the performance at low false accept rates as compared to performance at EER. We showcase that the proposed algorithm performs well even at a low false accept rate. 93

4-3 Illustrating the steps involved in the proposed face recognition algorithm. 94

4-4 Feature-richness distributions for two different videos. Some of the most feature-rich (values close to 1) and least feature-rich frames (values close to 0) are presented for illustration. We can see that the high fidelity frames are assigned a higher feature richness score and the poor frames which showcase artifacts such as occlusion and blur are assigned a low feature-richness score. Note that the total number of frames in the two videos is different. 95

4-5 Proposed deep learning architecture for facial representation: from input layer (image), two hidden layer representation are computed using SDAE encoding function. A joint representation is then obtained which combines the information from two SDAE encoding layers. Using joint representation as input, a DBM is used for computing a final feature vector. 101

4-6 Joint learning framework: features learned from the first and second levels of autoencoder, i.e., $\mathbf{f_1}$ and $\mathbf{f_2}$ are given as input to DBM to learn the joint representation $\mathbb{J}$. 102

4-7 ROC curves comparing the verification performance of the proposed algorithm with existing results reported on the YTF database webpage. 109

4-8 Summarizing the verification performance of the proposed algorithm and state-of-the-art algorithms on the YouTube Faces database. 110

4-9 Sample frames from the PaSC database: (a) random frames, (b) frontal frames, (c) most feature-rich frames, and (d) least feature-rich frames. 112

4-10 ROC curves comparing the verification performance of the proposed algorithm with frame selection approaches on the two databases. 113

5-1 We show that deep learning based OpenFace and VGG can be deceived even by image processing operations that mimic real world distortions. 120

5-2 Sample images representing the (b) grid based occlusion (Grids), (c) most significant bit based noise (xMSB), (d) forehead and brow occlusion (FHBO), (e) eye region occlusion (ERO), and (f) beard-like occlusion (Beard) distortions when applied to the (a) original images. (g) is the universal perturbed [148] images of PaSC and MEDS databases. 122

5-3 Bar graph showing the effect of perturbation on the VGG model. Verification accuracy is reported at 1% GAR. 127

5-4 Flow chart for the proposed detection and mitigation methodology. 128

5-5 Visualizing filter responses for selected layers from the VGG network when the input image is unaltered and affected by the grids and beard distortions. The first two rows present visualizations for conv3_2 and pool3 layers for the original input images respectively. The next two rows present visualizations for the same layers when the input images are distorted using adversarial perturbations. The propagation of the adversarial signal into the intermediate layer representations is the inspiration for our proposed detection and mitigation methodologies. 129

5-6 Visualizing the distribution of genuine (undistorted) and impostor (distorted) sample scores in the Canberra distance space. The scores are obtained by comparing the intermediate layer outputs for the sample with the layer mean obtained using the undistorted training data for a particular layer. In this illustration, we showcase the distributions for scores obtained using layers 3, 5, and 7 of the VGG network for the Multi-PIE database before normalization. 130

5-7 Summarizing the results of the proposed and existing detection algorithms on the PaSC and MEDS databases. 135

6-1 Illustrating a potential target for future work to improve face recognition further. The augmentation phase can fit in between the matching and decision phases and supplement face information with other cues that can aid overall recognition. . . . 139

List of Tables

1.1 Progression of state-of-the-art in face recognition algorithms in the past 10 years. . 7

2.1 Experimental protocol for both initial and extended experiments. 35

2.2 Identification accuracies (%) on the IIIT-D RGB-D face database and EURECOM database individually. The mean accuracy values are reported along with standard deviation. 43

2.3 Identification accuracy (%) for the extended gallery experiments. The mean accuracy values are reported along with standard deviation. 44

2.4 A detailed comparative analysis of the proposed algorithm with 3D-PCA, FPLBP, and Sparse approaches. T and F represent True and False respectively. True ground truth refers to genuine cases and false ground truth refers to the impostor cases. . . 44

2.5 Rank-1 identification accuracies on the 3D TEC [206] dataset. The results of other algorithms are presented as reported in [206]. 46

3.1 A literature review of feature-level fusion in biometrics. 50

3.2 Rank-1 identification accuracy (%) with individual features and their combination (SRC and GSRC) on the WVU and LEA databases. 69

3.3 Rank-1 identification accuracy (%) with fusion of multiple modalities and multiple features (SRC and GSRC) on the WVU and LEA databases. 72

3.4 The impact of feature characteristics on the performance of GSRC. ^{1} = original, ^{2} = normalized. 76

3.5 The impact of gallery size on the performance of GSRC. 76

3.6 Identification results at rank 5 on the SCFace database. Distance 1 is the furthest and distance 3 is the closest from the camera. 83

3.7 Identification results at rank 5 on both Eurecom and IIIT-D RGB-D databases. We use the existing protocol and partitions [66] to enable direct comparison. $*$ Results have been taken from [66]. 84

3.8 Rank-1 identification accuracy (%) with fusion of multiple modalities and multiple features (SRC, GSRC, and KGSRC) on the WVU database. 86

4.1 Review of selected papers on video face recognition that have shown results on the YTF and PaSC benchmark face video databases. Results marked unrestricted denote algorithms that have used external training data during training. The algorithms follow the standard experimental protocol during testing for both databases to facilitate comparison. ^{1}*denotes set based algorithms*, ^{2}*denotes sequence based algorithms*. 92

4.2 Details of the YTF and PaSC databases. 106

4.3 Verification rates on the PaSC database. Results of existing algorithms are reported from respective references. 111

4.4 Comparing the results of the proposed frame selection algorithm with existing image quality assessment algorithms and random frame selection. 114

4.5 Analyzing the performance of individual components of the proposed algorithm for face recognition. 115

4.6 GAR for cross database experiments at 0.01 FAR. 115

4.7 Comparing the verification accuracy of recent CNN based methods with the proposed algorithm. 116

5.1 Literature review of adversarial attack generation and detection algorithms. 121

5.2 Characteristics of the databases used for adversarial attack generation and detection. 124

5.3 Verification performance of existing face recognition algorithms in the presence of different distortions on the PaSC database. All values indicate genuine accept rate (%) at 1% false accept rate. 124

5.4 Verification performance of existing face recognition algorithms in the presence of different distortions on the MEDS database. All values indicate genuine accept rate (%) at 1% false accept rate. 125

5.5 Performance (accuracy %) of the proposed detection methodology (using LightCNN and VGG as the target networks) compared to two existing detection algorithms. Grids = grid based occlusion, xMSB = most significant bit based noise, FHBO = forehead and brow occlusion, ERO = eye region occlusion, and Beard = beard like occlusion. 133

5.6 Mitigation Results (GAR (%) at 1% FAR) on the MEDS and PaSC databases. . . . 135

Abbreviations

3D-TEC 3D-Twins Expression Challenge.

3DMM 3D Morphable face Models.

AAM Active Appearance Models.

ACNN Adaptive Convolutional Neural Network.

ADM Attributes based on Depth Map.

ADMM Alternating Direction Method of Multipliers.

BSC Block Sparse representation-based Classification.

CBFD Compact Binary Face Descriptor.

CMC Cumulative Match Characteristics.

CNN Convolutional Neural Network.

COTS Commerical Off-The-Shelf.

CR Collaborative Representation.

CRC-RLS CR based Classifier with Regularized Least Square.

CRSM Coupled Representation Similarity Metric.

CSLBP Center-Symmetric LBP.

CSML Cosine Similarity Metric Learning.

D-KSVD Discriminative K-SVD.

DBM Deep Boltzmann Machine.

DCP Dual-Cross Patterns.

DCS Discriminant Color Space.

DMSC Discriminative Multi-scale Sparse Coding.

DNN Deep Neural Network.

DPC Decision Pyramid Classifier.

DSNPE Discriminant Sparse Neighborhood Preserving Embedding.

DWT Discrete Wavelet Transform.

EER Equal Error Rate.

FAR False Accept Rate.

FOCUSS FOCally Underdetermined System Solver.

FPLBP Four-Patch LBP.

FRR False Reject Rate.

G-HFR Graphical representation based Heterogeneous Face Recognition.

GAR Genuine Accept Rate.

GMM Gaussian Mixture Models.

GPU Graphics Processing Unit.

GSRC Group Sparse Representation-based Classifier.

GTP Gabor Ternary Pattern.

HGPP Histogram of Gabor Phase Pattern.

HOG Histogram of Oriented Gradients.

ICP Iterative Closest Point.

IJB-A IARPA Janus Benchmark A.

JFL Joint Feature Learning.

K-SVD K-Singular Value Decomposition.

KGSRC Kernel Group Sparse Representation-based Classifier.

KSRC Kernel Sparse Representation based Classification.

LBP Local Binary Patterns.

LDA Linear Discriminant Analysis.

LDP Local Derivative Patterns.

LFW Labeled Faces in the Wild.

LPG Log Polar Gabor.

LTP Local Ternary Patterns.

MDML-DCP Multi-Directional Multi-Level DCP.

MEDS Multiple Encounter Dataset.

MKD Multi-Keypoint Descriptors.

MMC Maximum Margin Criterion.

MMV Multiple Measurement Vector.

NBIS NIST Biometric Image Software.

NIST National Institute of Standards and Technology.

NMR Nuclear-norm based Matrix Regression.

OMP Orthogonal Matching Pursuit.

PaSC Point and Shoot Challenge.

PCA Principal Component Analysis.

PDV Pixel Difference Vector.

PEP Probabilistic Elastic Part.

PHOG Pyramid Histogram of Oriented Gradients.

PMML Pairwise-constrained Multiple Metric Learning.

RBF Radial Basis Function.

RBM Restricted Boltzmann Machines.

RDF Random Decision Forest.

RDLRR Robust and Discriminative Low-Rank Representation.

RGB-D Red Green Blue-Depth.

RISE RGB-D Image descriptor based on Saliency and Entropy.

ROC Receiver Operating Characteristic.

ROI Region of Interest.

SAE Stacked Auto-Encoder.

SDAE Stacked Denoising Auto-Encoder.

SIFT Scale Invariant Feature Transform.

SMV Single Measurement Vector.

SNPE Sparse Neighborhood Preserving Embedding.

SOM Structured Ordinal Measure.

SPP Sparsity Preserving Projections.

SRC Sparse Representation-based Classifier.

SSEQ Spatial-Spectral Entropy-based Quality.

SURF Speeded Up Robust Features.

SVM Support Vector Machine.

TPLBP Three-Patch LBP.

UCLBP Uniform Circular LBP.

VASIR Video-based Automatic System for Iris Recognition.

VF VeriFinger.

WPCA Whitened Principal Component Analysis.

YTF YouTube Faces Database.

THIS PAGE INTENTIONALLY LEFT BLANK

Publications

Intellectual Property Disclosures

1. **G. Goswami**, M.Vatsa, N. Ratha, R. Singh, S. Pankanti, Identifying Artificial Artifacts in Input Data to Detect Adversarial Attacks, US Patent Application filed, 2018.

2. **G. Goswami**, M.Vatsa, N. Ratha, R. Singh, S. Pankanti, Mitigating False Recognition of Altered Inputs in Convolutional Neural Networks, US Patent Application filed, 2018.

JOURNALS

1. **G. Goswami**, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa, Detecting and Mitigating Adversarial Perturbations for Robust Face Recognition, International Journal of Computer Vision, **Impact Factor: 11.541** (Submitted after revision).

2. P. Chhokra, A. Chowdhury, **G. Goswami**, M. Vatsa, and R. Singh, Unconstrained Kinect Video Face Database, Information Fusion, 2018, Volume 44, Pages 113-125, 2018, **Impact Factor: 6.639**.

3. **G. Goswami**, M. Vatsa, and R. Singh, Video Face Verification via Learned Representation on Feature-Rich Frames, IEEE Transactions on Information Forensics and Security, Volume 12(7), Pages 1686-1698, 2017, **Impact Factor: 5.824**.

4. P. Mittal, A. Jain, **G. Goswami**, M. Vatsa, and R. Singh, Composite sketch recognition using saliency and attribute feedback, Information Fusion, Volume 33, Pages 86-99, 2017, **Impact Factor: 6.639**.

5. A. Sankaran, **G. Goswami**, M. Vatsa, R. Singh, and A. Majumdar, Class Sparsity Signature based Restricted Boltzmann Machines, Pattern Recognition, Volume 61, Pages 674-685, 2016, **Impact Factor: 3.962**.

6. **G. Goswami**, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, Group Sparse Representation based Classification for Multi-feature Multimodal Biometrics, Information Fusion, Volume 32(B), Pages 3-12, 2016, **Impact Factor: 6.639**.

7. **G. Goswami**, M. Vatsa, and R. Singh, RGB-D Face Recognition with Texture and Attribute Features, IEEE Transactions on Information Forensics and Security, Volume 9(10), Pages 1629-1640, October 2014, **Impact Factor: 5.824**.

8. **G. Goswami**, B. Powell, M. Vatsa, R. Singh, and A. Noore, FR-CAPTCHA: CAPTCHA Based on Recognizing Human Faces. PLoS ONE, Volume 9, 2014, **Impact Factor: 2.766**.

9. B. Powell, **G. Goswami**, M. Vatsa, R. Singh, and A. Noore, fgCAPTCHA: Genetically Optimized Face Image CAPTCHA, IEEE Access, Volume 2, Pages 473-484, 2014, **Impact Factor: 3.557**.

10. **G. Goswami**, B. Powell, M. Vatsa, R. Singh, and A. Noore, FaceDCAPTCHA: Face detection based color image CAPTCHA, Future Generation Computer Systems, Volume 31, Pages 59-68, 2014, **Impact Factor: 4.639**.

PEER REVIEWED CONFERENCE ARTICLES

1. **G. Goswami**, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks, Thirty-Second AAAI Conference on Artificial Intelligence, 2018. **(Oral Presentation)**

2. **G. Goswami**, R. Singh, M. Vatsa, A. Majumdar, Kernel Group Sparse Representation based Classifier for Multimodal Biometrics, 30th International Joint Conference on Neural Networks, 2017.

3. **G. Goswami**, N. Ratha, M. Vatsa, and R. Singh, Improving Classifier Fusion via Pool Adjacent Violators Normalization, 23rd International Conference on Pattern Recognition, Pages 1011-1016, 2016.

4. B. Powell, **G. Goswami**, M. Vatsa, R. Singh, and A. Noore, A Multibiometrics-based CAPTCHA for Improved Online Security, 6th IEEE International Conference on Biometrics: Theory, Applications and Systems, 2016.

5. R. Bhardwaj, **G. Goswami**, R. Singh and M. Vatsa, Harnessing Social Context for Improved Face Recognition, IAPR International Conference on Biometrics, 2015.

6. A. Jain, P. Mittal, **G. Goswami**, M. Vatsa and R. Singh, Person Identification at a Distance via Ocular Biometrics, IEEE International Conference on Identity, Security and Behavior Analysis, 2015.

7. **G. Goswami**, R. Bhardwaj, M. Vatsa, and R. Singh, MDLFace: Memorability augmented deep learning for video face recognition, IEEE/IAPR International Joint Conference on Biometrics, 2014. **(Oral Presentation)**

8. P. Mittal, A. Jain, **G. Goswami**, M. Vatsa, and R. Singh, Recognizing Composite Sketches with Digital Face Images via SSD Dictionary, IEEE/IAPR International Joint Conference on Biometrics, 2014.

9. **G. Goswami**, S. Bharadwaj, M. Vatsa, and R. Singh, On RGB-D Face Recognition using Kinect, 6th IEEE International Conference on Biometrics: Theory, Applications and Systems, 2013 **(Received the Best Poster Award)**.

10. **G. Goswami**, R. Singh, M. Vatsa, B. M. Powell, and A. Noore, Face Recognition CAPTCHA, 5th IEEE International Conference on Biometrics: Theory, Applications and Systems, Pages 412-417, 2012.

BOOK CHAPTERS

1. **G. Goswami**, M. Vatsa, and R. Singh, Face Recognition with RGB-D images using Kinect, in Face Recognition across the Imaging Spectrum, Springer International Publishing, 2016,

pp. 281-303.

2. T.I. Dhamecha, **G. Goswami**, R. Singh, and M. Vatsa, On Frame Selection for Video Face Recognition, in Advances in Face Detection and Facial Image Analysis, Springer International Publishing, 2016, pp. 279-297.

3. **G. Goswami**, R. Singh, and M. Vatsa, Automated Spam Detection in Short Text Messages, in Machine Intelligence and Signal Processing, Springer India, 2016, pp. 85-98.

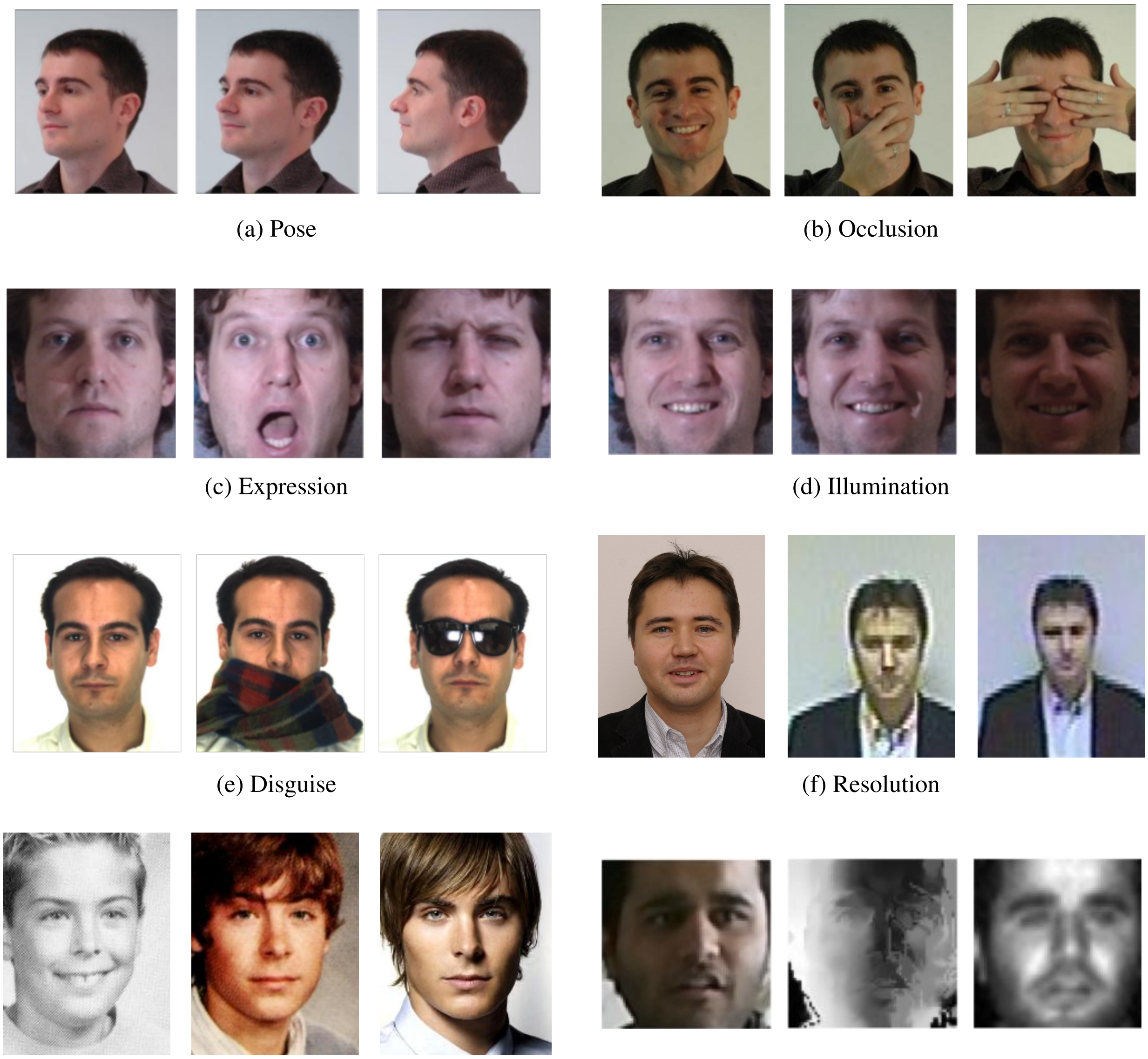
Chapter 1

Introduction

Face recognition is a task that humans perform every day swiftly, accurately, and repeatedly. However, when automated algorithms are used to perform the same task, it becomes a challenging research problem which has received proportionate attention in the literature. Even after decades of active research in this area, face recognition algorithms struggle to achieve the consistently accurate performance of the human mind for familiar faces. Even though certain algorithms have been demonstrated to be better than human face recognition in particular restricted scenarios [127], there is no single algorithm that can provide robust and consistent recognition accuracy in all real world situations.

Face is one of the most easily accessible biometric modality that does not require special acquisition procedures and cooperation of the subject, which are reasons that make it useful for a wide variety of applications ranging from automated photo tagging on social media platforms to critical applications such as border control and surveillance forensics. Therefore, even though fingerprint and iris technologies are more accurate and mature [106], they cannot completely replace the need for face biometrics. However, human faces also demonstrate high inter-class and intra-class similarities due to the same overall structure, look-alikes, pose, illumination, expression, occlusion, and disguise. There are further cross-view challenges [150] for automated algorithms such as low resolution [188] and cross spectral variations [58] that are required to be addressed. Figure 1-1 provides examples for each of these covariates or challenges of face recognition.

A typical face recognition algorithm involves multiple stages. First, an input image is processed to detect and crop the face region. The detected region of interest is then aligned on the basis of



(a) Pose

(b) Occlusion

(c) Expression

(d) Illumination

(e) Disguise

(f) Resolution

(g) Age

(h) Spectrum

Figure 1-1: Illustrating different covariates that deter the accuracy of face recognition algorithms. Images are taken from the AR face database [141], the CMU Multi-PIE database [69], the SCFace database [68], the Large Age Gap (LAG) database [18], and the KaspAROV database [30].

facial landmarks such as the eyes, nose, and mouth to adjust for scale, shift, and orientation and then fed into a feature extractor [153]. A feature extractor converts the information contained in the image to a numeric vector called as the feature vector which is a mathematically comparable representation of the face. Ideally, representations extracted from faces of different individuals are sufficiently distinguishable from representations extracted from different face images of the same individual. Depending on the algorithm there might be a single feature extractor or many of them which can later be combined at a feature level by concatenation or other techniques. There are two modes in which face recognition can be performed that determine the final stages of the recognition pipeline. If the mode is one-to-one matching, also termed as face verification, then the features extracted from one face image (probe) are matched directly with those extracted from another (gallery). This matching might be performed using a previously learned classifier such as a Support Vector Machine (SVM) [195] or a simple L-2 distance metric. The value of this comparison is used to obtain the match score that can be further compared with a threshold to decide whether the two faces belong to the same individual or not. If the mode is one-to-many matching, also termed as face identification, then the features extracted from the input face (probe image) are compared to gallery face images belonging to all the individuals enrolled in a database and individuals are sorted according to the value of the comparison metric. The sorted list of individuals is called a ranked list and the identity of the individual in the probe face image is predicted as the identity with the highest match score. Both the modes have real world applications and it is a common practice for algorithms to address verification directly and then perform identification in verification mode by making multiple one-to-one comparisons to emulate the one-to-many matching.

The accuracy of face recognition relies heavily on the feature extraction process. As shown in Figure 1-1, there are several factors that can alter the appearance of the same face significantly in two different images and make feature extraction challenging. Over time, researchers have proposed novel methodologies to impart resilience against particular covariates at a time. However, such algorithms have had limited scalability and success in a truly unconstrained environment where multiple covariates are present in a single image. Therefore, there has been a paradigm shift in the way face recognition algorithms are evaluated wherein recent databases try to capture an unconstrained environment by imposing less restrictions on the capture process. Images/videos of subjects are captured while performing one or the other activity which ensures that multiple variations are present simultaneously in almost every data point. To handle data variations from an unconstrained capture process, there are two possible approaches: (1) at the input level; enhance the input data itself in terms of the quantity, quality, and modality, and (2) at the algorithm level; improve the feature extraction methodology. In this dissertation, we explore both directions while keeping other important factors such as practical scalability and feasibility in consideration.

Improving feature extraction by enhancing the input itself can be achieved by increasing the dimensionality, combining different modalities, adding different views, and adding temporal information. Two-dimensional face images in the visible domain have been used as the input to traditional face recognition algorithms. While this is a compact and convenient way to capture face data, it is severely limited in the amount of information that it can contain. It has been deduced in the literature that capturing a 3D object (such as a face) in a two-dimensional format has inherent problems. On the other hand, capturing a 3D image requires sensors that are still not commonplace and fall in the category of specialized equipment which has an associated premium in terms of setup and cost. In order to leverage the additional information provided by the depth capture of a face while limiting the increase in cost to create practically feasible face recognition methodologies, we explore the trade-off offered by consumer level RGB-D capture devices such as Kinect. While such devices do not capture *true* 3D images, they provide a pseudo 3D representation at a low cost. These RGB-D face images present their own set of challenges: the depth maps are of relatively low quality and the alignment with the color image is not perfect. Adding temporal information and different views can also be accomplished by utilizing a face video instead of still images. In a single video a face can be captured with many different pose, illumination, and expression variations allowing for a more *complete* feature extraction since the algorithm gets visibility on how the same face changes under the effect of these covariates. There are challenges associated with using video data as well: processing requirements are high and not all the frames contain useful information. Therefore, a mechanism to select only the most relevant subset of frames is important to make the incorporation of video data more useful and feasible.

Improving feature extraction by improving the methodology can also manifest itself in various ways: extracting different types of features, ensuring robustness of features to various covariates, and fusing multiple features into an ensemble. While initial research in facial features relied on handcrafted representations, deep learning based methodologies have become popular to learn the best way to represent input data using data-driven training. While the obtained representations are highly discriminative and offer large accuracy improvements, processing large amounts of training data into a very deep network is computationally intensive and increases the dependency on data volume. It is also important to not just improve the performance of algorithms on selected databases and controlled data but ensure that the improvements are sustained in a real setting for unseen data. A particular challenge in achieving this is the existence of adversarial input. The generation of adversarial inputs has received special attention in the literature since it exposes the singularities present in the existing methodologies for automatic face recognition. It is possible to completely change the decisions made by an algorithm by introducing minor changes to the data which may not be immediately visible to a human observer. Also, while some defense mechanisms are proposed to handle such input, they are not highly generalizable and most of them require modifications to the network and/or the training process and re-training. An ideal defense mechanism would not suffer from the same limitations.

Many face recognition algorithms involve information fusion at either the feature, score, or decision level. This allows for various individual feature descriptors that are individually capable of handling only a limited set of covariates to produce a more robust overall output. However, a lot of the information contained in a feature is lost when the fusion takes place at the score or decision level. Therefore, feature-level fusion is an important area of focus. Particularly when the complimentary information is available. In this dissertation, we explore four research directions, viz, (i) RGB-D face recognition with texture and attribute features, (ii) group sparse representation based classifier for feature-level fusion, (iii) feature richness and joint deep representation based framework for video face recognition, and (iv) defense against adversarial attacks, and attempt to utilize their advantages while addressing the associated challenges. Figure 1-2 provides a broad overview of this dissertation and summarizes its contributions.

1.1 Face Recognition: Progress from 2007 to 2018

Most of the research in face recognition literature has been focused on improving the alignment, feature extraction, and matching aspects of the pipeline. In this section, we explore how the feature extraction methodologies have evolved with time.

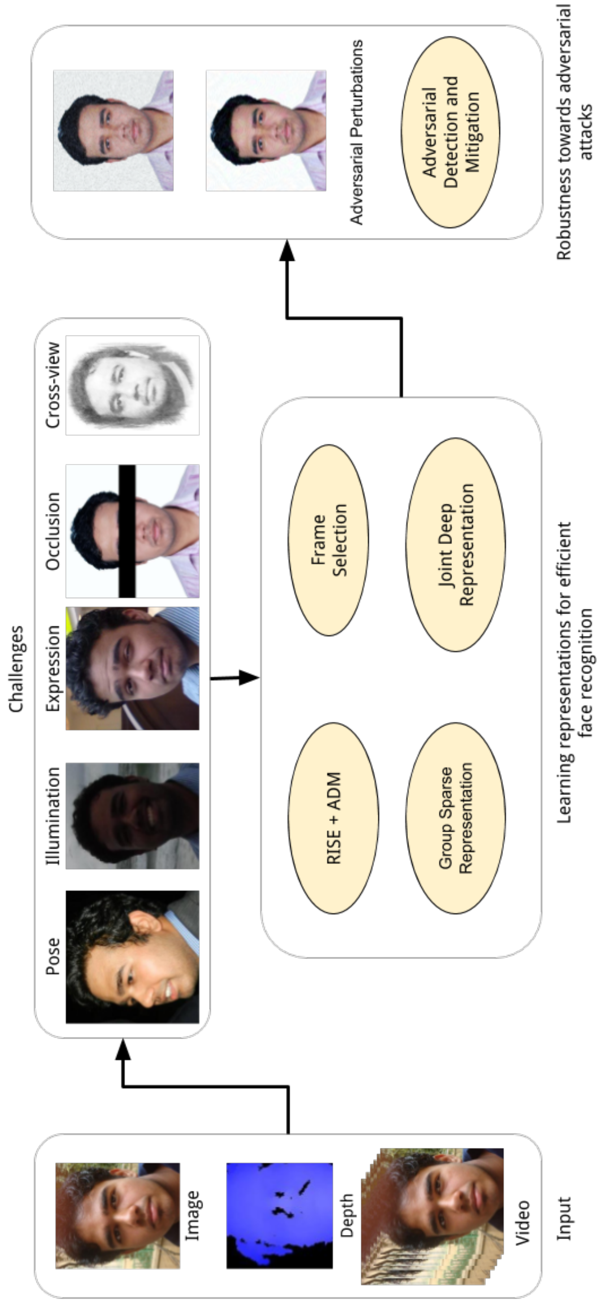

Input
Image
Depth
Video
Challenges
Pose
Illumination
Expression
Occlusion
Cross-view
RISE + ADM
Group Sparse Representation
Frame Selection
Joint Deep Representation
Learning representations for efficient face recognition
Adversarial Perturbations
Adversarial Detection and Mitigation
Robustness towards adversarial attacks

Figure 1-2: Illustrating the scope of this dissertation. Using multiple modes of input such as single face image, depth image in addition to face image, and face video, we solve the challenges in extracting efficient features pertaining to the covariates of pose, illumination, expression, cross-view, and occlusion. We also propose adversarial detection and mitigation techniques to ensure the robustness of the representations in a real world scenario.

Table 1.1: Progression of state-of-the-art in face recognition algorithms in the past 10 years.

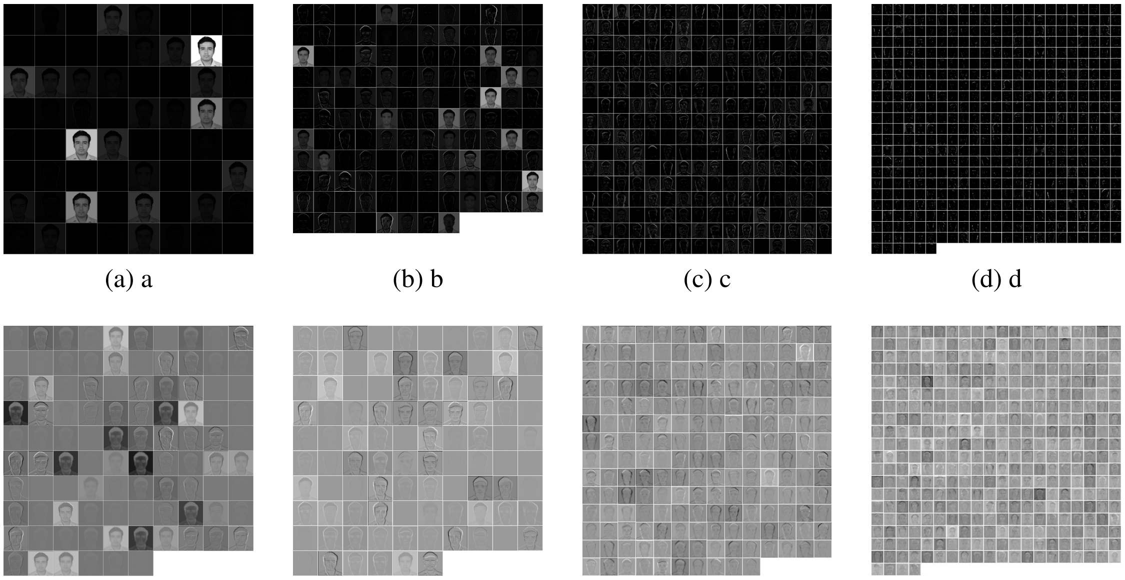
Year	Labeled Faces in the Wild	YouTube Faces Database
2008	Ensemble of LBP, Gabor, TPLBP, and FPLBP features [215]	
2009	Information Theoretic Metric Learning + LBP, SIFT, TPBLP, and FPLBP [199]	
2010	Cosine Similarity Metric Learning + LBP, Gabor, and intensity [152]	
2011	Large scale feature-search with neuromorphic feature representations [33]	Matched Background Similarity + LBP, CSLBP, and FPLBP [114]
2012	Distance Metric Learning with Eigenvalue Optimization + SIFT [231]	
2013	SIFT + Fisher Vectors + Joint-Metric Similarity Learning [187]	Sparse Coding + Whitened PCA + Pairwise constrained Multiple Metric [234] Learning
2014	Deep Convolutional Neural Networks (DeepFace, DeepID) [192, 222]	
2015	FaceNet: Deep Convolutional Neural Network [183]	
2016	LBPNet: Local Binary Pattern Network (LBP + CNN) [219]	Discriminative 3D Morphable Models with a very Deep Convolutional Neural Network [3]
2017-2018	Probabilistic Elastic Part Model + LBP and SIFT [117]	Feature-richness based frame selection with SDAE + DBM based joint feature representation [67]

1.1.1 2007-2013: Fusion of Handcrafted Features and Distance Metric Learning

In the initial years, all of these efforts were based around proposing new hand-crafted features and fusing them together to further increase the accuracy and robustness of face recognition approaches. In 2007, Zhang *et al.* [236] proposed the Histogram of Gabor Phase Pattern (HGPP) descriptor that encodes quadrant-bit codes based on global and local Gabor transformations and encodes Gabor phase information as opposed to magnitude. In the same year, Huang *et al.* [83] introduced the first widely used challenging unconstrained benchmark database for face recognition with a set protocol that researchers could use to report results and benchmark their algorithms. In 2008, Wolf *et al.* [215] obtained the best performance of 78.47% classification accuracy on the Labeled Faces in the Wild (LFW) database using a combination of hand-crafted texture features such as Local Binary Patterns (LBP), Gabor, Three-Patch LBP (TPLBP), and Four-Patch LBP (FPLBP) to obtain multiple match scores for each pair of images and using these scores as a feature vector with a SVM classifier. In 2009, Taigman *et al.* [199] used a One-Shot Similarity measure that utilized labeled data outside of the restricted protocol and combined it with information theoretic metric learning to learn the similarity metric. They extracted Scale Invariant Feature Transform (SIFT), LBP, TPLBP, and FPLBP features and used a similar approach to [215] to create feature vectors for use with a SVM classifier to perform verification. Zhang *et al.* [235] proposed another local texture descriptor termed as the Local Derivative Patterns (LDP) that encodes directional pattern features based on local derivative variations. The *n*th-order LDP is proposed to encode the (n-1)th-order local derivative direction variations as opposed to the first-order patterns encoded by LBP. In the same year, Wright *et al.* [217] framed the problem of face recognition using sparse signal representation theory calling it the Sparse Representation-based Classifier (SRC). They showed that by harnessing sparsity, the choice of features becomes less important than whether the features are large enough and correct computation of the sparse representation. They demonstrate that using this formulation, unconventional features such as downsampled images and random projections perform comparably to conventional features such as eigenfaces, as long as the dimension of the feature space surpasses a certain threshold. This research sparked a long list of further research inspired by introducing sparsity in features for face recognition.

In 2010, Nyugen *et al.* [152] improved the state-of-the-art for face verification using Gabor and LBP textures and raw intensity values of the input images as features with Cosine Similarity Metric Learning (CSML) obtaining a verification accuracy of 88.0%. Tan *et al.* [200] proposed Local Ternary Patterns (LTP), a generalization of LBP that was more discriminant and less sensitive to noise in uniform regions. They combined it with Kernel PCA feature extraction with two additional feature sources: Gabor wavelets and LBP. Their approach obtained encouraging results on various databases used for testing face recognition algorithm under varying illumination conditions. Cao *et al.* [21] presented an approach named learning-based descriptor that encodes the micro-structures of the face by a new learning-based encoding method. They utilized unsupervised learning techniques to learn an encoder from the training examples followed by PCA to get the face descriptor with reduced dimensionality. Their approach achieved comparable to state-of-the-art performance on the LFW database. Qiao *et al.* [164] proposed an unsupervised dimensionality reduction algorithm called Sparsity Preserving Projections (SPP) that focused not on preserving local neighborhood information but the sparse reconstructive relationship of the data. They achieved this using a L1 regularization-related optimization criterion to obtain a projection that is invariant to rotation, scale, and translation, and contains discriminating information with automatically chosen neighborhoods. Zhang *et al.* [240] proposed the Discriminative K-SVD (D-KSVD) algorithm to extend the K-SVD algorithm by incorporating the classification error as part of the optimization criterion, thus allowing the performance of a linear classifier and the representational power of the dictionary being optimized at the same time. The D-KSVD algorithm computes an overcomplete dictionary and solves for the classifier using a procedure derived from the K-SVD algorithm. They have also shown that the learned dictionary and classifier are indeed better for sparse-representation-based recognition.

In 2011, Wolf *et al.* [114] introduced another benchmark database called the YouTube Faces Database (YTF) that presented an even more challenging video face recognition problem for comparative evaluation of unconstrained face recognition algorithms. When we consider the algorithms that were state-of-the-art on these databases over the past decade, we can see how the best performing algorithms have evolved. An overview of this progression is presented in Table 1.1. In the same year, Zhang *et al.* [238] explored the use of Collaborative Representation (CR) in the SRC algorithm and claimed that it is the CR and not the L1-norm sparsity that contributes to its success in face recognition. Based on their observations, they also proposed an alternative algorithm, namely CR based Classifier with Regularized Least Square (CRC-RLS). Cox *et al.* [33] offered an alternative to the texture features that had been popular till then, proposing the first form of learning features from data. They proposed a large-scale feature search approach that chose the best candidates for the face verification task from a collection of randomly generated multilayer neuromorphic representations. On the YTF, the very first algorithms described were still based on texture features such as LBP, Center-Symmetric LBP (CSLBP), and FPLBP used in conjunction with a matched background similarity measure to account for the set based nature of face video matching [114]. In 2012, Deng *et al.* [38] proposed an extension of the SRC to the single gallery face recognition problem that utilized samples from other classes to compute the sparse representation for a probe image. Gui *et al.* [71] proposed a new sparse subspace learning algorithm called Discriminant Sparse Neighborhood Preserving Embedding (DSNPE) combining discriminant information into Sparse Neighborhood Preserving Embedding (SNPE) that utilizes the global discriminant structures using a MMC added to the objective function. Ma *et al.* [134] proposed a discriminative low-rank dictionary learning algorithm for sparse representation motivated by low-rank matrix recovery using an objective function with sparse coefficients, class discrimination, and rank minimization. Distance metric learning remained a key focus, with Ying *et al.* [231] pushing the boundaries of state-of-the-art with a novel eigenvalue optimization framework based on the Mahalanobis metric. Again, the features used in this research were hand-crafted texture features namely SIFT, LBP, and TPLBP. In 2013, Liao *et al.* [121] presented an alignment-free partial face recognition approach based on Multi-Keypoint Descriptors (MKD), where the descriptor size for a given face image depends on the image content enabling any image to be sparsely represented within a large dictionary of proposed Gabor Ternary Pattern (GTP) descriptors. Li *et al.* [115] presented the first algorithm that utilized a low resolution 3D sensor for improving face recognition under challenging conditions using additional depth data. They exploited facial symmetry to fill in holes in the obtained 3D point cloud and then sparse approximated the depth and texture face data using separate dictionaries learned from training data and observed high recognition rates by using the depth data. Simonyan *et al.* [187] demonstrated that Fisher vectors on densely sampled SIFT features could achieve state-of-the-art face verification performance when combined with a joint-metric similarity learning that encoded the difference between a low-rank inner product and a low-rank Mahalanobis distance between the extracted features. The focus was again on the metric learning aspect of the algorithm to improve the performance. In the same year, Cui *et al.* [234] presented an approach to compute features for video matching using spatial blocks comprised of frames and combined sparse codes obtained from members of each block by sum pooling. They used a Whitened Principal Component Analysis (WPCA) for refining the feature dimension and also proposed a distance metric learning method called Pairwise-constrained Multiple Metric Learning (PMML) to effectively integrate the collection of descriptors. State-of-the-art algorithms available by the end of 2013 were able to achieve up to 93% verification accuracy on the LFW database.



(a) a

(b) b

(c) c

(d) d

(e) e

(f) f

(g) g

(h) h

Figure 1-3: Illustrating how a deep network extracts facial features. Images (a) to (d) represent the output of progressively deeper convolution layers from the VGG deep network. Images (e) to (h) represent outputs from convolution layers of equivalently increasing depth from the LightCNN network. As we can see, the first layer of output consists of basic patterns primarily simulating simple edge detection that continue to evolve into more complex combinations of edges and texture in the later layers of the network. The learned features are optimized to obtain the best recognition performance and therefore capture information that focuses on preserving the differentiability of different faces.

1.1.2 2014-2018: A Focus on Deep Learning

In recent years, research in face recognition has shifted from hand-crafted features to automatically learned representations using deep networks. An illustration of how such a representation framework extracts features is presented in Figure 1-3. The year of 2014 marked a transition in the face recognition literature from the prevailing trend of using metric learning with an ensemble of hand-crafted texture descriptors to data-driven learned representations. The boundaries of state-of-the-art were pushed by multiple contributions all based around deep convolutional neural networks combined with the usage of large unrelated training databases in the unrestricted protocols for both the benchmark databases. Deep learning algorithms such as DeepFace [222] and DeepID [192] achieved the best performance in 2014 reaching 97.4% verification performance on the LFW database. In 2015, FaceNet [183] combined distance metric learning and representation learning in one unified embedding learning framework that furthered the state-of-the-art to 99.6% on the LFW database. Ding *et al.* [41] proposed a comprehensive deep learning framework to jointly learn face representation using multimodal information. They utilized an array of specially designed Convolutional Neural Network (CNN)s and a three-layer Stacked Auto-Encoder (SAE). While the CNNs extract facial features from the multi-modal data, the SAE performs dimension reduction on the concatenated high-dimensional feature vector. Ding *et al.* [45] also proposed a novel face identification framework capable of addressing a multitude of pose variations by converting the problem into a partial frontal face recognition problem. They then deploy a patch-based face representation scheme with transformative dictionaries for improved recognition. Sun *et al.* [193] presented a deep convolutional network termed DeepID2+ for face recognition using supervised signals in early convolution layers to further the state-of-the-art on the LFW [83] and YTF [114] benchmark databases. They identified sparsity, selectiveness, and robustness as the key properties for the high performance of the network. Schroff *et al.* [183] proposed FaceNet, that maps inputs from the face space to a Euclidean space where distances between points directly relate to face similarity. Using these FaceNet features, they formulate face recognition problems in terms of standard problems such as clustering and finding the nearest neighbors. They utilized CNNs to optimize the extracted feature embedding and report results better than DeepID2+ on the same benchmark databases. Sun *et al.* [191] then proposed two very deep neural network architectures named DeepID3. They based this off of the stacked convolution and inception layers proposed in VGG net [160] and GoogLeNet [196] to adapt them for face recognition. They then reinforce them with the supervised signals that distinguished DeepID2+. Lu *et al.* [129] presented a new Joint Feature Learning (JFL) algorithm to learn features from raw pixels for face recognition. They proposed an unsupervised feature learning method to compute hierarchical features using different feature dictionaries to represent them on a facial region basis. Using spatial pooling and stacking of these features, they further improved the recognition accuracies. Lu *et al.* [130] also proposed CBFD based face representations by extracting Pixel Difference Vector (PDV)s in localized facial patches in a neighborhood. They then learn a mapping from these PDVs into low-dimensional binary vectors in an unsupervised manner while maximizing the variance of all binary codes, minimizing the loss between the original real-value and learned binary codes, and ensuring even distribution of binary codes at each learned bin. The final features are obtained by pooling the codes into a histogram. Lu *et al.* [131] also proposed a new algorithm for image-set-based face recognition that utilized statistics information to compute the features for each face set. They combine it with a localized multikernel metric learning algorithm that enables learning feature-specific distance metrics in the kernel spaces for each statistic. Liu *et al.* [122] illustrated a two-stage approach that combined a multi-patch deep CNN and deep metric learning to extract compact features for face recognition without sacrificing performance. Parkhi *et al.* [160] presented a deep CNN using 16 convolution layers that is trained on an augmented database of 2.6 million face images pertaining to 2,625 individuals to improve the state-of-the-art in face recognition on multiple benchmark databases showing another case of data-driven features providing high accuracy.

This trend of learning features for improved performance continued into 2016. Xi *et al.* [219] presented an unsupervised deep learning based methodology combining the topology of convolutional neural networks by replacing the learnable filters in the convolution layers with LBP and PCA inspired filters to achieve competitive results with the state-of-the-art algorithms without needing as much training data. In the same year, Tran *et al.* [201] regressed the shape and texture parameters to create efficient 3D Morphable face Models (3DMM) using a convolutional neural network to obtain results comparable to state-of-the-art deep learning algorithms while generating interpretable representations in the form of 3D face shapes. Ding *et al.* [40] proposed a novel algorithm to exploit the first derivative of Gaussian operator to handle illumination variations and then computed Dual-Cross Patterns (DCP) features at both the holistic and component levels. Through experiments on multiple databases they conclude that their proposed feature descriptor termed Multi-Directional Multi-Level DCP (MDML-DCP) outperforms other existing hand-crafted feature descriptors in both face identification and verification tasks. Masi *et al.* [143] explored if the daunting task of collecting a lot of face images is absolutely essential for face recognition algorithms. They proposed face specific data augmentation techniques to generate multiple training samples from a single face sample in a training database. They reported that the performances of their approach match that of algorithms trained on millions of faces. We report similar results in Chapter 4 where we observe that the proposed approach is able to perform comparably to the state-of-the-art methods even using limited training data and even without the need for data augmentation along with an analysis of how the algorithm performs with increased data. Zhang *et al.* [241] proposed an extension to CNNs which can compute the optimal structure of the network, termed as Adaptive Convolutional Neural Network (ACNN). They initialized the network by a one-branch structure and the average error and recognition rate of the training samples are set to control the expansion of the structure of CNN. They first extend the network globally until the average error criterion is met and then expanded locally to meet the recognition rate criterion. Chen *et al.* [25] presented an algorithm for unconstrained face verification based on deep convolutional features and evaluated it on the IARPA Janus Benchmark A (IJB-A) [101] and LFW [83] databases. They used the CASIA-WebFace dataset [229] for training their CNN. AbdAlmageed *et al.* [3] proposed a method for face recognition using multiple pose-aware deep learning models. They processed a face image by multiple pose-specific deep CNN models to generate corresponding pose-specific features. They then utilized 3D-rendering techniques to generate multiple poses of the same face image. They introduced robustness to pose variations by using this ensemble of features. Sun *et al.* [194] proposed an improvement to the high face recognition performance of deep CNNs by introducing sparsity in neural connections. They learned these sparse ConvNets in an iterative manner where sparsity is introduced in each layer iteratively and re-training of the model is performed using the weights of the previous iterations as the starting points. They reported that instantiating a sparse ConvNet by basing it off of a pre-trained dense model is critical for extracting good features for face recognition. Wen *et al.* [214] proposed a new supervision signal for training CNNs instead of the traditionally used softmax to improve the representative power of the learned features for face recognition, which they call center loss. The center loss simultaneously learned a center for a particular class and penalized features for individual samples that deviated too far from their class center. They also proved that the center loss function is easily optimized for CNN training and used it in conjunction with softmax loss to train CNNs with inter-class separability and intra-class similarity as key objectives. They observed improvements on benchmark databases when compared to state-of-the-art methods including other CNN based approaches.

In 2017, Li *et al.*, [117] proposed an approach that again utilizes LBP and SIFT to extract features from densely sampled multi-scale image patches and encoded their location in combination with the feature information itself. They capture the spatial-appearance distribution for all face images using Gaussian Mixture Models (GMM), terming it the Probabilistic Elastic Part (PEP) model. The PEP representation is built by concatenating the feature descriptors by each component in the GMM. In chapter 4, we discuss a feature-richness based frame selection and joint representation approach that achieves state-of-the-art results on the YTF with a deep architecture that comprises of fewer parameters and can perform well without needing large amounts of training data. In summary, we have observed an evolution of the face recognition literature shifting attention from ensembles of hand-crafted texture descriptors to distance metric learning and now to data-driven learned features primarily using deep neural networks. Yang *et al.* [226] explored the use of regression analysis for face recognition. They have focused on addressing the occlusion and illumination covariates by utilizing low-rank structural information. They accomplish this with a two-dimensional image-matrix-based error model, termed as the Nuclear-norm based Matrix Regression (NMR). Their proposed approach computes the minimal nuclear norm of feature error image as a metric and the Alternating Direction Method of Multipliers (ADMM) is deployed to compute the regression coefficients. They observed better results compared to other regression based approaches. Peng *et al.* [162] proposed a novel Graphical representation based Heterogeneous Face Recognition (G-HFR) method that is based on Markov networks which extract features from heterogeneous image patches separately, and also encodes the spatial compatibility of neighboring image patches. They also proposed a Coupled Representation Similarity Metric (CRSM) to compare the graphical representations. Hayat *et al.* [75] discussed that a single representation may not be sufficient for an image set and instead argue for retaining the images of a set in their original form and using their proposed extensions of popular simple binary classifiers for multi-class image recognition for both objects and faces. They showed that their approach can perform well even with training few binary classifiers on limited training data. He *et al.* [76] presented a data driven Structured Ordinal Measure (SOM) based method that captures both ordinal filters and structured ordinal features geared towards face recognition in videos. Unlike hand-crafted ordinal measures, their proposed approach learns the filters and further enforces low-rankness and optimality of the ordinal matrix for accurate classification. They also combine these features with deep features that are jointly utilized for improved stability of the obtained feature encoding. Gao *et al.* [55] focused on imparting robustness towards occlusion in face recognition. They have capitalized on the low-rankness of the feature and the occlusion-induced error images to create the structure information preserving Robust and Discriminative Low-Rank Representation (RDLRR). Huang *et al.* [84] proposed a facial landmark-based multi-scale LBP feature descriptor that can address pose and expression variations. They then fused LBP and Gabor features at kernel-level to encode both facial texture and shape information while automatically learning the optimal values for the parameters involved using their proposed optimization algorithm. Pei *et al.* [161] presented a novel face recognition algorithm termed Decision Pyramid Classifier (DPC) directed towards solving the challenge of single sample per person, especially addressing pose and expression covariates and occlusion to some extent. Their proposed DPC algorithm is nonparametric and extracts features from a multitude of non-overlapping local facial regions from both training and test images and constructs a decision pyramid to determine the identity of an unseen probe image. Yu *et al.* [232] also proposed another solution for the single sample per person problem utilizing a Discriminative Multi-scale Sparse Coding (DMSC) model for robustness towards occlusion in particular. They compute an occlusion variation model using disjoint training data and create a dictionary and use it to classify pixels as outliers and obtain a sparse and efficient feature encoding. Gao *et al.* [57] presented yet another method for this problem using semi-supervised sparse representation-based classification. They characterize the face recognition problem in terms of a gallery dictionary that may contain one or more gallery faces and a variation dictionary that contains covariate information, which they term as *nuisances*. The algorithm primarily consists of utilizing the variation dictionary to encode the covariates with sparsity and estimating prototype face images using the gallery dictionary and a Gaussian mixture model, with mixed labeled and unlabeled samples in a semi-supervised manner. In 2018, the research has continued primarily with deep learning based approaches. Masi *et al.* [142] have proposed using an ensemble of pose specific CNNs trained to tackle extreme pose variations to create a pose aware model for unconstrained face recognition. Wen *et al.* [213] have proposed a unified framework to encapsulate both global and local features by providing a splitting behavior for certain intermediate features which can be handled in different branches. They do this within the same CNN instead of requiring multiple CNNs that individually learn to process different levels of features based on the type of facial patches that they are trained on. Lu *et al.* [133] train three deep networks to address the problem of mismatching low resolution in face recognition based on surveillance videos. They train one of these networks on a mix of high and low resolution data and two resolution-specific networks that together extract relevant discriminative features and perform better recognition in a resolution limited scenario. Sparse representation based techniques are also being explored in the literature. Fan *et al.* [50] have proposed a kernel sparse representation based approach wherein they utilize approximately symmetrical face images to artificially increase training data and a coordinate descent approach to solve the sparsity constrained optimization problem. Liu *et al.* [125] also propose a variant of kernel sparse representation where they formulate the sparse optimization problem as a weighted L1 minimization and utilize multi scale retinex to generate the similarity matrix while computing the sparse representation of the test sample.

1.2 Research Contributions

As discussed in the previous section, there exists a substantial amount of literature and prior research in the area of improving automated face recognition algorithms. However, face recognition in the real world is still far from a solved problem since there exist many gaps in the capabilities of the methodologies proposed in the existing literature as compared to the complexities of the problem. Therefore, as part of this dissertation, we have attempted to fill some of these gaps and progressively improve various aspects of representations for face recognition. There can be different modalities in which a face image can be captured and expressed which influence the kind of data available to any representation technique that operates with it. These are illustrated in Figure 1-4. We focus on the three types of representation modalities indicated in the figure as well as

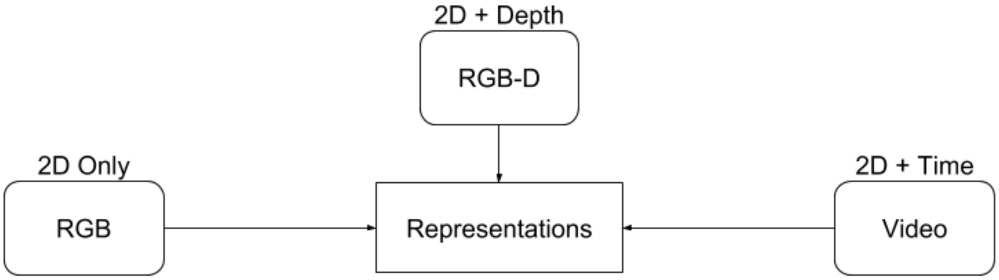


Figure 1-4: Features can be extracted from various modalities of face images. While a traditional RGB (2D image) provides only color and texture information, it may be combined with depth information to obtain a RGB-D based representation, or may be combined with a time dimension (by capturing the same object across multiple frames separated by acquisition time) to obtain a video based representation.

address the robustness of deep learning based approaches against adversarial attacks. The major contributions of this dissertation are:

- **RGB-D Face Recognition with Texture and Attribute Features**: Although methodologies using more than just 2D images for face recognition exist, they require high cost specialized 3D sensors. In order to make it more feasible to use depth-informed algorithms to improve face recognition performance, we introduce a novel algorithm using RGB-D images. The proposed algorithm computes a descriptor based on the entropy of RGB-D faces along with the saliency feature obtained from a 2D face. Geometric facial attributes are also extracted from the depth image and face recognition is performed by fusing both the descriptor and attribute match scores. IIIT-D RGB-D face database of 106 individuals is prepared and shared with the research community to promote further research in this area. A detailed experimental protocol along with train-test splits are also shared to encourage other researchers to report comparative results. This research is published in IEEE BTAS 2013 [61] and in the IEEE Transactions on Information Forensics and Security 2017 [67].

- **Group Sparse Representation based Classification**: There are many existing feature extraction techniques proposed in the literature with many representations capturing uncorrelated and complementary information about face images. In this case, significant performance improvement can be expected if multiple features can be leveraged in unison to perform recognition. However, fusion techniques in the literature have primarily combined features at the score or decision level where much of the information contained in the complete feature vectors is lost and merely considered at the aggregate level. Therefore, we propose the Group Sparse Representation-based Classifier (GSRC) which removes the requirement for a separate feature-level fusion mechanism and integrates multi-feature seamlessly into classification. By considering each feature source without the use of concatenation or feature reduction, the classification algorithm can utilize different feature spaces to make an optimal decision. The performance of the proposed GSRC classifier is evaluated with multiple feature sets and biometric modalities on two publicly available databases. We also propose a kernel extension to the GSRC which enables multiple features to be processed in a higher dimensional space where they are more separable, without substantially increasing computational costs. The proposed Kernel Group Sparse Representation-based Classifier (KGSRC) algorithm selects the ideal kernel to use along with its parameters automatically as part of the training process. We evaluate the proposed algorithm on three challenging biometric problems namely, RGB-D face recognition, cross distance face recognition, and multimodal biometrics to showcase its efficacy. This research is published in Information Fusion 2015 [63] and IJCNN 2017 [65].

- **Feature-Richness and Joint Representation for Video Face Recognition**: Videos have been explored in the literature as a modality for face information as they are a means to obtain temporal information and thereby improve the fidelity of the extracted features. Not all the information contained in a video may be relevant for obtaining robust features and therefore frame selection is an important aspect of video face recognition. While most of the existing literature performs frame selection randomly, based on selected pose, or based on a quality metric based threshold, we propose a novel face verification algorithm which starts with selecting feature-rich frames from a video sequence using discrete wavelet transform and entropy computation. Frame selection is followed by representation learning based feature extraction where three contributions are presented: (i) deep learning architecture which is a combination of Stacked Denoising Auto-Encoder (SDAE) and Deep Boltzmann Machine (DBM), (ii) formulation for joint representation in an autoencoder, and (iii) updating the loss function of DBM by including sparse and low rank regularization. Finally, a multilayer neural network is used as classifier to obtain the verification decision. The results are demonstrated on two publicly available databases, YTF and Point and Shoot Challenge (PaSC) database. Experimental analysis suggests that (i) the proposed feature-richness based frame selection offers noticeable and consistent performance improvement compared to frontal only frames, random frames, or frame selection using perceptual no-reference image quality measures, and (ii) joint feature learning in SDAE and sparse and low rank regularization in DBM helps in improving face verification performance. On the benchmark PaSC database, the algorithm yields the verification accuracy of over 97% at 1% false accept rate whereas on the YTF, over 95% verification accuracy is observed at equal error rate. This research is published in the IEEE Transactions on Information Forensics and Security 2017 [67].

- **Evaluating and Addressing the Robustness of Deep Representations for Face Recognition against Adversaries**: Deep learning based methodologies have become increasingly more popular with state-of-the-art results on challenging databases. However, there are challenges in translating the performance achieved by such a network in the presence of adversaries. In order to be able to confidently deploy a deep learning based solution to real world problems, we attempt to unravel three aspects related to the robustness of Deep Neural Network (DNN)s for face recognition: (i) assessing the impact of deep architectures for face recognition in terms of vulnerabilities to attacks inspired by commonly observed distortions in the real world; (ii) detecting the singularities by characterizing abnormal filter response behavior in the hidden layers of deep networks; and (iii) making corrections to the processing pipeline to alleviate the problem. Our experimental evaluation using two open-source DNN-based face recognition networks, OpenFace and VGG, and two publicly available databases (Multiple Encounter Dataset (MEDS) and PaSC) demonstrates that the performance of deep learning based face recognition algorithms can suffer greatly in the presence of such distortions. The proposed SVM-based method is able to detect the attacks almost 100% of the time by suitably designing a classifier using the response of the hidden layers in the network. Finally, we present several effective countermeasures to mitigate the impact of adversarial attacks and improve the overall robustness of DNN-based face recognition. This research is published in AAAI 2018 [64].

Chapter 2

RGB-D Face Recognition with Texture and Attribute Features

2.1 Introduction

Face recognition with 2D images is a challenging problem especially in the presence of covariates such as pose, illumination, and expression. These covariates introduce high degree of variation in two 2D images of the same person thereby reducing the performance of recognition algorithms [96]. Therefore, it is desirable to perform face recognition using a representation which is less susceptible to such distortions. While 2D images are not robust to these covariates, 3D images offer a comparatively resilient representation of a face. 3D images can capture more information about a face, thus enabling higher preservation of facial detail under varying conditions. 3D face recognition has been explored in literature and several algorithms have been developed [20, 46, 180]. While it is advantageous to utilize 3D images for face recognition, the high cost of specialized 3D sensors limits their usage in large scale applications.

With advancements in sensor technology, low cost sensors have been developed that provide (pseudo) 3D information in the form of RGB-D images. As shown in Figure 2-1, an RGB-D image consists of a 2D color image (RGB) along with a depth map (D). RGB image provides the texture and appearance information whereas depth map provides the distance of each pixel from the sensor. The depth map is a characterization of the geometry of the face with grayscale values representing the distance of each point from the sensor. While a RGB-D image does not provide

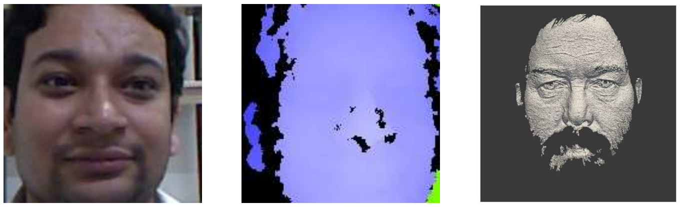


Figure 2-1: Different modes of capture: (a) RGB image, (b) depth map captured using Kinect, and (c) Range image from 3D TEC dataset [206] obtained using a 3D scanner.

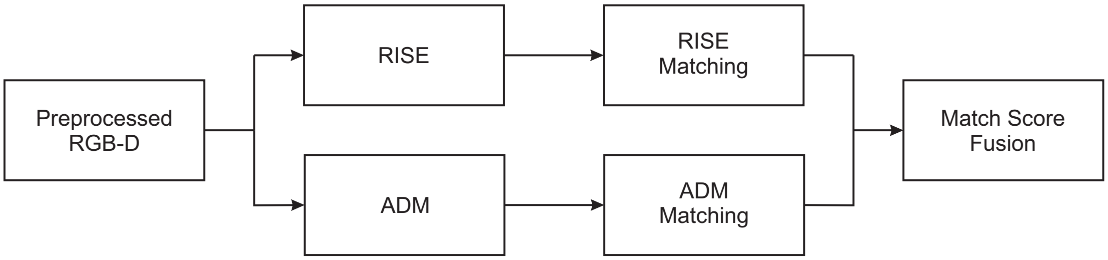


Figure 2-2: Illustrating the steps involved in the proposed RGB-D face recognition algorithm.

highly accurate 3D information, it captures more information compared to a 2D image alone.

An RGB-D image captured using consumer devices such as Kinect is fundamentally different from a 3D image captured using range sensors due to the manner in which they capture the target. Kinect captures RGB-D image by utilizing an infrared laser projector combined with a monochrome CMOS sensor. 3D sensors on the other hand utilize specialized high quality sensors to obtain accurate range and texture image. 3D face recognition approaches utilize techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to characterize a 3D face model. Some approaches also utilize facial landmarks identified in a 3D face model to extract local features. However, 3D face recognition algorithms generally rely on accurate 3D data. Since the depth map returned by RGB-D Kinect sensor is not as precise as a 3D sensor and contains noise in the form of holes and spikes, existing 3D face recognition approaches may not be directly applied to RGB-D images. While RGB-D images have been used for several computer vision tasks such as object tracking, face detection, gender recognition, and robot vision [49, 77, 78, 81, 89, 167], there exists relatively limited work in face recognition. Li et al. [115] proposed a face recognition framework based on RGB-D images. The RGB-D face image obtained from Kinect is cropped using the nose tip which is reliably detectable via the depth map. The face is then transformed into a canonical frontal representation and pose correction is performed using a reference face model. The missing data is filled by symmetric filling which utilizes the symmetry of human faces to approximate one side of the face with corresponding points from the other side. Smooth resampling is then performed to account for holes and spikes. The image is converted using Discriminant Color Space (DCS) transform [210, 225] and the three channels are stacked into one augmented vector. This vector and the depth map are individually matched via Sparse Representation-based Classifier (SRC) [217] and the scores are combined. Experimental results indicate that using both depth and color information yields around 6% higher identification accuracy compared to color image based algorithms. Segundo et al. [184] proposed a continuous face authentication algorithm which utilizes Kinect as the RGB-D sensor. The detected face image is aligned to an average face image using the Iterative Closest Point (ICP) algorithm [10] and a Region of Interest (ROI) is extracted. The ROI is then characterized using Histogram of Oriented Gradients (HOG) approach and utilized for matching with stored user template for authentication. Kinect also has its own algorithm for face recognition, the details of which are not publicly available.

While there are few algorithms that utilize RGB-D images obtained from consumer devices for face recognition, this research presents a different perspective. As mentioned previously, the depth maps obtained using Kinect are noisy and of low resolution. Therefore, instead of using the depth information to generate a 3D face model for recognition, we utilize noise tolerant features for extracting discriminatory information. We propose a novel face recognition algorithm that operates on a combination of entropy and saliency features extracted from the RGB image and depth entropy features extracted from the depth map. The proposed algorithm also utilizes geometric attributes of the human face to extract geometric features. These geometric features are utilized in conjunction with the entropy and saliency features to perform RGB-D face recognition. The key contributions of this research are:

- A novel algorithm is developed that uses both texture (oriented gradient descriptor based on saliency and entropy features) and geometric attribute features for identifying RGB-D faces.

- IIIT-D RGB-D face database of 106 individuals is prepared and shared with the research community to promote further research in this area. A detailed experimental protocol along with train-test splits are also shared to encourage other researchers to report comparative results.

2.2 Proposed RGB-D Face Recognition Algorithm

The steps involved in the proposed algorithm are shown in Figure 2-2. The algorithm is comprised of four major steps: (a) preprocessing, (b) computing textute descriptor from both color image and depth map using entropy, saliency, and HOG [35], (c) extracting geometric facial features from depth map, and (d) combining texture and geometric features for classification. These steps are explained in the following subsections.

2.2.1 Preprocessing

First, an automatic face detector (Viola-Jones face detector) is applied on the RGB image to obtain the face region. Any other face detection framework may also be applied [169]. The corresponding region is also extracted from the depth map to crop the face region in depth space. While texture feature descriptor does not require image size normalization, the images are resized to 100×100 to compute depth features. Depth map is then preprocessed to remove noise (holes and spikes). Depth map of a face is divided into 25×25 blocks and each block is examined for existence of holes and spikes. Depth values identified as the hole/spike are rectified using linear interpolation, i.e. assigned the average value of their 3×3 neighborhood.

2.2.2 RISE: RGB-D Image descriptor based on Saliency and Entropy

The motivation of the proposed *RGB-D Image descriptor based on Saliency and Entropy (RISE)* algorithm lies in the nature of the RGB-D images produced by Kinect. Specifically, as shown in Figure 2-3, depth information obtained from Kinect has high inter-class similarity and may not be directly useful for face recognition. It is our assertion that 3D reconstruction based approaches may not be optimal in this scenario. However, due to low intra-class variability, depth data obtained from Kinect can be utilized to increase robustness towards covariates such as expression and pose after relevant processing/feature extraction. On the other hand, 2D color images can provide inter-class differentiability which depth data lacks. Since the color images contain visible texture properties of a face and the depth maps contain facial geometry, it is important to utilize both RGB and depth data for feature extraction and classification. As shown in Figure 2-4, four entropy maps corresponding to both RGB and depth information and a visual saliency map of the

RGB image are computed. The HOG descriptor [35] is then used to extract features from these five entropy/saliency maps. The concatenation of five HOG descriptors provides the texture feature descriptor which is used as input to the trained Random Decision Forest (RDF) classifier to obtain the match score.

Entropy and Saliency

Entropy is defined as the measure of uncertainty in a random variable [178]. Similarly, the entropy of an image characterizes the variance in the grayscale levels in a local neighborhood. The entropy H of an image neighborhood $\mathbf{x}$ is given by Equation 2.1,

$$H(\mathbf{x}) = -\sum_{i=1}^{n} p(x_i) log_b p(x_i) \tag{2.1}$$

where $p(x_i)$ is the value of the probability mass function for x_i. In the case of images, $p(x_i)$ signifies the probability that grayscale x_i appears in the neighborhood and n is the total number of possible grayscale values, i.e., 255. If $\mathbf{x}$ is a $M_H \times N_H$ neighborhood then

$$p(x_i) = \frac{n_{x_i}}{M_H \times N_H} \tag{2.2}$$

Here, n_{x_i} denotes the number of pixels in the neighborhood with value x_i. $M_H \times N_H$ is the total number of pixels in the neighborhood. By controlling the size of neighborhood, entropy computation can be performed at a fine or coarse level. In the current research, the neighborhood size for entropy map computation is fixed at 5×5 and RGB input images are converted to grayscale. The visual entropy map of an image is a characteristic of its texture and can be used to extract meaningful information from an image. Examples of entropy and depth entropy maps are presented in Figure 2-4. The absolute values of the depth entropy map do not vary abruptly in adjacent regions except in special regions such as near the eye sockets, nose tip, mouth, and chin. The local entropy of an image neighborhood measures the amount of randomness in texture (in local region). Higher local entropy represents higher prominency in that region and therefore, it can be viewed as a texture feature map that encodes the uniqueness of the face image locally and allows for a robust feature extraction.

Apart from entropy, we also utilize *visual saliency* of the RGB image to compute useful facial

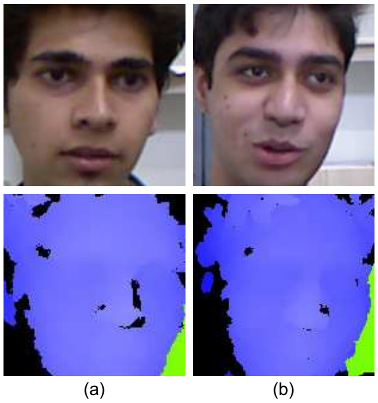


Figure 2-3: RGB-D images of two subjects illustrating the inter-class similarities of RGB images and depth maps.

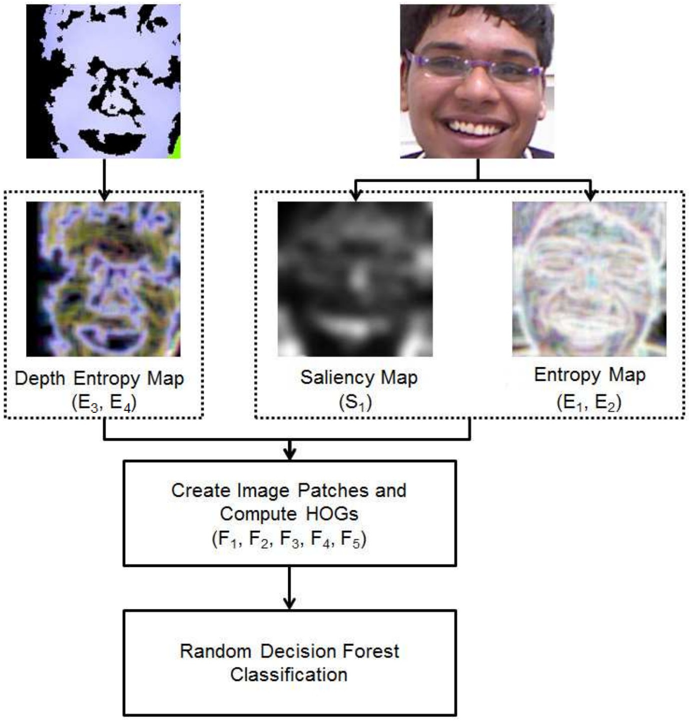


Figure 2-4: Illustrating the steps of the proposed RISE algorithm.

information. It measures the attractiveness of local regions based on the viewer's visual attention [39]. The distribution of visual attention across the entire image is termed as visual saliency map of the image. There are several approaches to compute the visual saliency map of an image. This research utilizes the approach proposed by Itti et al. [90]. Let the image be represented as an intensity function which maps a set of co-ordinates (x, y) to intensity values. The approach preprocesses a color image to normalize the color channels and de-couple hue from intensity. After normalization, center-surround differences are utilized to yield the feature maps [90]. 42 feature maps are extracted from the image in accordance with the visual cortex processing in mammals. Six of these maps are computed for intensity, 12 for color, and 24 for orientation across multiple scales. Intensity and orientation feature maps are denoted by I and O respectively. The color feature maps are represented by RG and BY which are created to account for color double opponency in the human primary visual cortex [48]. Based on these maps, the saliency map of the image is computed by cumulating the individual feature maps obtained at different scales to one common scale ($= 4$) of the saliency map. This is achieved after inhibiting the feature maps which are globally homogeneous and promoting the maps which comprise of few unique activation spots (global maxima) via a normalization function $N(\cdot)$. The feature maps for color, intensity and orientation are combined in separate groups to create three feature maps C_{final}, I_{final}, and O_{final} corresponding to color, intensity, and orientation respectively.

$$C_{final} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [N(RG(c,s)) + N(BY(c,s))] \tag{2.3}$$

$$I_{final} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c,s)) \tag{2.4}$$

$$O_{final} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} N\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c,s,\theta)) \right) \tag{2.5}$$

Here, c and s denote the center and surround scales respectively and the $\bigoplus$ operator denotes across-scale addition which is defined to consist of reduction of each map to the common scale and point-wise addition [90]. These maps are then combined into the final visual saliency map S

according to equation 2.6:

$$S = \frac{1}{3}[N(C_{final}) + N(I_{final}) + N(O_{final})] \tag{2.6}$$

Figure 2-4 presents an example of the visual saliency map, S, of an input face image. It models the image regions with high feature activation in accordance with the visual processing that occurs in the visual cortex of mammals. It is observed that gradient orientations of this saliency map provide discriminative information which aids in improving the recognition performance, specifically in reducing the intra-class discrepancies. Therefore, orientation histogram of the saliency map of a color image (obtained using HOG approach) is utilized as an additional feature. It is to be noted that saliency is computed only for RGB image and not depth map because the depth map lacks salient information and therefore, the saliency of depth map does not provide discriminating information.

Extracting Entropy Map and Visual Saliency Map

Let the input RGB-D image be denoted as $[I_{rgb}(x, y), I_d(x, y)]$, where $I_{rgb}(x, y)$ is the RGB image and $I_d(x, y)$ is the depth map, both of size $M \times N$. Let both of these be defined over the same set of (x, y) points such that $x \in [1, M]$ and $y \in [1, N]$. Let $H(I_j)$ denote the visual entropy map of image I_j. Here, I_j can be the depth map or the RGB image or a small part of these images. Two image patches are extracted for both I_{rgb} and I_d. Two patches, P_1 of size $\frac{M}{2} \times \frac{N}{2}$ centered at $[\frac{M}{2}, \frac{N}{2}]$, and P_2 of size $\frac{3M}{4} \times \frac{3N}{4}$ centered at $[\frac{M}{2}, \frac{N}{2}]$ are extracted from I_{rgb}. Similarly, two patches P_3 and P_4 are extracted from I_d. Four entropy maps $E_1 - E_4$ are computed for patches $P_1 - P_4$ using Equation 2.7:

$$E_i = H(P_i),\ \ where\, i \in [1, 4] \tag{2.7}$$

E_1, E_2 represent the entropy of the color image (I_{rgb}) and E_3, E_4 represent the depth entropy maps.

The proposed RISE algorithm also extracts visual saliency map S_1 of the color image I_{rgb} using Equation 2.8.

$$S_1(x, y) = S(I_{rgb}(x, y)\forall(x \in [1, M], y \in [1, N])) \tag{2.8}$$

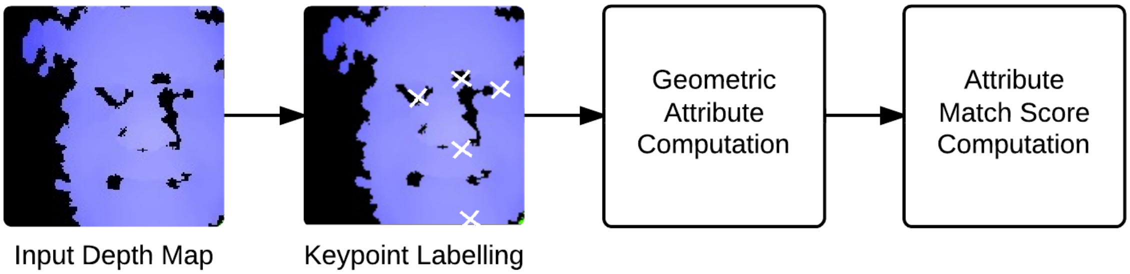

Input Depth Map
Keypoint Labelling
Geometric Attribute Computation
Attribute Match Score Computation

Figure 2-5: Steps involved in the proposed ADM approach.

Extracting Features using HOG

HOG [35] descriptor produces the histogram of a given image in which pixels are binned according to the magnitude and direction of their gradients. HOG has been successfully used as a feature and texture descriptor in many applications related to object detection, recognition, and other computer vision problems [32, 52, 209]. HOG of an entropy map or saliency map encodes the gradient direction and magnitude of the image variances in a fixed length feature vector. The information contained in the entropy/saliency map can therefore be represented compactly with a HOG histogram. Further, histogram based feature encoding enables non-rigid matching of the entropy/saliency characteristics which may not be possible otherwise.

In the proposed RISE algorithm, HOG is applied on the entropy and saliency maps. The entropy maps are extracted from patches P_i which allows capturing multiple granularities of the input image. Let $D(\cdot)$ denote the HOG histogram; the proposed algorithm computes HOG of entropy maps using the following equation:

$$F_i = D(E_i), \, where \, i \in [1, 4] \tag{2.9}$$

Here, F_1 represents the HOG of entropy map E_1 defined over patch P_1 and F_2 represents the HOG of entropy map E_2 defined over patch P_2 of I_{rgb}. Similarly, F_3 and F_4 represent the HOG of entropy maps E_3 and E_4 defined over patches P_3 and P_4 of I_d respectively. F_1 and F_2 capture traditional texture information but instead of directly using visual information, entropy maps are used to make the descriptor robust against intra-class variations. F_3 and F_4 capture the depth information embedded in the RGB-D image.

Next, HOG descriptor of visual saliency map, S_1 is computed using Equation 2.10. The final descriptor F is created using an ordered concatenation of the five HOG histograms as shown in Equation 2.11.

$$F_5 = D(S_1(I_{rgb})) \tag{2.10}$$

$$F = [F_1, F_2, F_3, F_4, F_5] \tag{2.11}$$

Concatenation is used to facilitate training by reducing five vectors to a single feature vector. Since each HOG vector is small, the resulting concatenated vector has a small size which helps in reducing the computational requirement. The feature vector F is provided as input to a multi-class classifier.

Classification

To establish the identity of a given probe, a multi-class classifier such as Nearest Neighbor (NN), Random Decision Forests (RDFs) [80], and Support Vector Machines (SVM) can be used. However, the classifier should be robust for large number of classes, computationally inexpensive during probe identification, and accurate. Among several choices, RDFs being an ensemble of classifiers, can produce non-linear decision boundaries and handle the multi-class classification. RDFs are also robust towards outliers compared to the Nearest Neighbor algorithm, since every tree in the forest is only trained with a small subset of data. Therefore, the probability of an entire collection of trees making an incorrect decision due to a few outlier data points is very low. Moreover, as per the experimental results in the preliminary research, RDF is found to perform better than NN [61]. Other classifiers such as SVM requires significant more training data per class. Therefore, in this research, RDF is used for classification. In RDF training, the number of trees in the forest and the fraction of training data used to train an individual tree control the generalizability of the forest. These parameters are obtained using the training samples and a grid search. Here, each feature descriptor is a data point and the subject identification number is the class label, therefore, the number of classes is equal to the number of subjects. The trained RDF is then used for probe identification. A probe feature vector is input to the trained RDF which provides a probabilistic match score for each class. This match score denotes the probability with which the feature vector belongs to a particular class. To summarize, the RISE algorithm is presented in Algorithm 1.

```
Data: Preprocessed RGB-D image, $I_{rgb}$, denotes the color image and $I_d$ denotes the depth
      map
Result: The RISE descriptor for the given RGB-D image $F$
for $i \leftarrow 1$ to $2$ do
    $E_i$ = Entropy map of patch $P_i$ of $grayscale(I_{rgb})$;
end
for $i \leftarrow 3$ to $4$ do
    $E_i$ = Entropy map of patch $P_i$ of $I_d$;
end
$S$ = Saliency map of $I_{rgb}$;
$E_5$ = Entropy map of $S$;
for $i \leftarrow 1$ to $5$ do
    $F_i$ = HOG of $E_i$;
    $F$ = Concatenation of $H_i$;
end
```

Algorithm 1: The RISE algorithm

2.2.3 ADM: Attributes based on Depth Map

Attributes based methodologies have been applied successfully in image retrieval [102, 112] and face verification [111]. In RGB-D face recognition, it can be an additional useful feature. However, instead of qualitative or descriptive attributes such as gender, age, and complexion, the proposed Attributes based on Depth Map (ADM) algorithm extracts geometric attributes. Multiple geometric attributes can be utilized to describe a face such as the distances between various key facial features such as eyes, nose, and chin. By exploiting the uniform nature of a human face, key facial landmarks can be located and utilized to extract geometric attributes that can be used for face recognition in conjunction with the entropy and saliency features. An overview of the ADM approach is illustrated in Figure 2-5. The ADM approach consists of the following steps.

Keypoint Labeling

To extract geometric attributes, first a few facial key points are located with the help of depth map. The points such as nose tip, eye sockets, and chin can be extracted by using a "rule template". In a detected face depth map, the nose tip is closest point from the sensor, the two eye sockets are always located above the nose tip and at a higher distance than their local surrounding regions (due to cheek bones and eyebrows being at a lesser distance), the chin can be detected as the closest point to the sensor below the nose tip. Utilizing these key points, some other landmarks such as the nose bridge and eyebrow coordinates can also be located. By using a standard set of landmarks for all faces, a consistent way to compute geometric measurements of the face is possible.

Geometric Attribute Computation

To obtain the geometric attributes, various distances between these landmark points are computed: inter-eye distance, eye to nose bridge distance, nose bridge to nose tip distance, nose tip to chin distance, nose bridge to chin distance, chin to eye distance, eyebrow length, nose tip distance to both ends of both eyebrows, and overall length of the face. Since the measured value of these parameters may vary across pose and expression, multiple gallery images are utilized to extract the facial features. Attributes are computed individually for each gallery image and the distances are averaged. In this manner, a consistent set of attributes is computed for a subject. These contribute towards the attribute feature vector for the RGB-D face image.

Attribute Match Score Computation

The attributes for a probe are computed similar to gallery images. Once the attributes are computed for a probe, the match score Φ is computed for each subject in the gallery using Equation 2.12.

$$\Phi = \sum_{i=1}^{N} w_i \times (A_i - a_i)^2 \tag{2.12}$$

Here, A_i and a_i are the i^{th} attributes of the probe image and the gallery image respectively. w_i is the weight of the i^{th} attribute and N is the total number of attributes. w_i is used to assign different weights to different attributes depending upon how reliably they can be computed. In this research, w_i is optimized using grid search for efficient identification performance on the training dataset. After computation, the match scores from each subject can be utilized for identification. However, in the proposed approach it is combined with the match score obtained by RISE algorithm for taking the final decision.

2.2.4 Combining RISE and ADM

The match scores obtained by RISE and ADM algorithms can be combined in various ways. In this research, we explore two types of fusion:

Match Score Level Fusion

Many match score level fusion techniques have been proposed in the literature to utilize the different characteristics of the match scores being combined for improving the performance [109, 110]. In this research, we choose to go with a simple weighted sum rule [175] so that we can evaluate the baseline performance gains just by combining RISE and ADM. Choosing a specialized fusion technique may further increase the performance depending on other factors. Let Φ_{RISE} be the match score obtained using the RISE approach and Φ_{ADM} be the match score obtained by the ADM approach. The fused match score Φ_{fused} is computed as,

$$\Phi_{final} = w_{RISE} \times \Phi_{RISE} + w_{ADM} \times \Phi_{ADM} \tag{2.13}$$

where w_{RISE} and w_{ADM} are the weights assigned to the RISE and ADM match scores respectively.

Rank Level Fusion

Many rank level fusion algorithms have been proposed in the literature to utilize different characteristics of the ranks being combined for improved performance [107]. In this research, we choose to go with a simple weighted Borda count [175] so that we can evaluate the baseline performance gains just by combining RISE and ADM at the rank level. Choosing a specialized fusion technique may further increase the performance depending on other factors. Weighted Borda count allocates a score to a subject depending on its rank in both the ranked lists and then creates a new ranked list for identification based on these scores. The ranked list of subjects is created using both RISE and ADM match scores individually. These ranked lists are then combined by computing a new match score for each subject based on these ranked lists according to Equation 2.14.

$$Rf_{subj} = \sum_{i=RISE,ADM} \sum_{j=1}^{R_{max}} \begin{cases} w_i(R_{max} - j) & if\ R_{ij} = rank(subj) \\ 0 & otherwise \end{cases} \tag{2.14}$$

Here, R_{max} denotes the maximum (worst) possible rank value. w_{RISE} and w_{ADM} denote the weights for RISE and ADM respectively. Similarly, R_{RISE} and R_{ADM} denote the ranked lists of RISE and ADM respectively. The weights w_{RISE} and w_{ADM} can be used to control the number of points that the ranked lists of RISE and ADM can provide to the subject. $R_{ij} = rank(subj)$ signifies the condition that the $subject$ has rank j in the i^{th} ranked list.

2.3 Experimental Results

The performance of the proposed approach is analyzed via two types of experiments. First, the experiments are conducted on the IIIT-D RGB-D dataset to analyze the performance of the proposed approach with various combinations of constituent components and their parameters. Thereafter, the performance is compared with existing 2D and 3D approaches on an extended dataset.

2.3.1 Database and Experimental Protocol

There are a few existing RGB-D databases in literature. The EURECOM [89] database has 936 images pertaining to 52 subjects and the images are captured in two sessions with variations in pose, illumination, view, and occlusion. The VAP RGB-D [78] face database contains 153 images pertaining to 31 individuals. The dataset has 51 images for each individual with variations in pose. However, both of these datasets contain images pertaining to a relatively small number of individuals. To evaluate the performance of face recognition, a larger dataset is preferable. Therefore, the IIIT-D RGB-D database[1] is prepared for the experiments. This database contains 4605 RGB-D images pertaining to 106 individuals captured in two sessions using Kinect sensor

[1] The database and ground truth information is available via https://research.iiitd.edu.in/groups/iab/rgbd.html

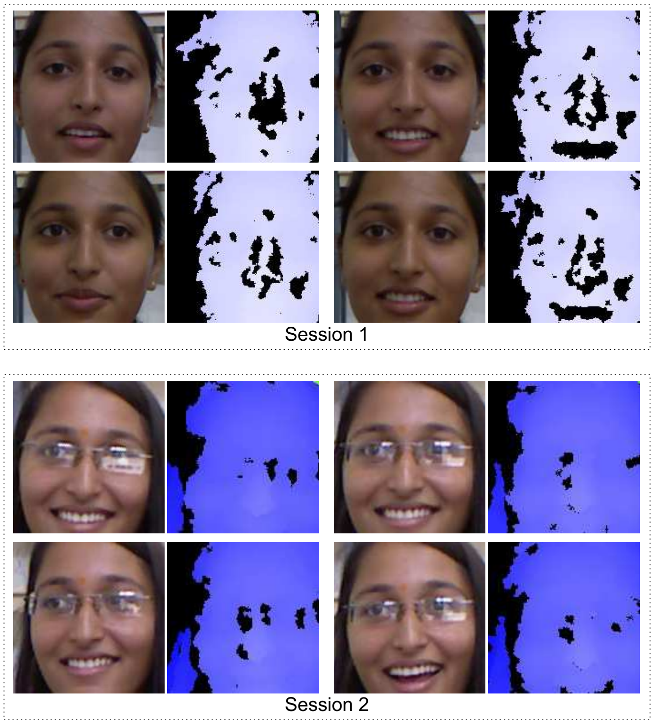


Figure 2-6: Sample images of a subject in two sessions from the IIIT-D RGB-D database.

Table 2.1: Experimental protocol for both initial and extended experiments.

Experiment	**Database**	**No. of Images**	**No. of Subjects**	
			Training	**Testing**
Experiment 1	IIIT-D RGB-D	4605	42	64
Experiment 2	IIIT-D RGB-D + VAP + EURECOM	5694	75	114

and OpenNI SDK. The resolution of both the color image and the depth map is 640×480. The number of images per individual is variable with a minimum of 11 images and a maximum of 254 images. In this database, the images are captured in normal illumination with variations in pose and expression (in some cases, there are variations due to eye-glasses as well). Some sample images for a subject in the IIIT-D database are presented in Figure 2-6. Using these three datasets, two types of experiments are performed. The initial experiments are performed on the IIIT-D RGB-D dataset to analyze the component-wise performance of the proposed RISE approach as well as to study the impact of weights and gallery size on the identification performance. Thereafter, the IIIT-D RGB-D dataset is merged with the EURECOM [89] and VAP [78] datasets to create an extended dataset of 189 individuals. The extended dataset is used to compare the performance of the proposed algorithm with existing 2D and 3D approaches.

The experimental protocol for each experiment is detailed below:

- **Experiment 1:** 40% of the IIIT-D Kinect RGB-D database is used for training and validation. The training dataset is utilized to compute the weights involved in ADM approach, RDF classifier parameters, and weights for fusion. Note that RDF classifier is separately trained for the initial and extended experiments by utilizing the respective training datasets. After training and parameter optimization, the remaining 60% dataset (unseen subjects) is used for testing. The results are computed with five times random subsampling. In each iteration of the subsampling, the subjects chosen for training/testing as well as the gallery images selected for each subject are randomly selected. Gallery size is fixed at four images per subject.

- **Experiment 2:** The extended database of 189 subjects is used for this experiment. Images pertaining to 40% individuals from the extended database are used for training and the remaining 60% unseen subjects are used for testing. To create the complete subject list for the extended dataset, the subjects are randomly subsampled within the three datasets according to 40/60 partitioning and then merged together to form one extended training/testing partition. Therefore, the extended training dataset has proportionate (40%) representation from each of the three datasets. The number of images available per individual varies across the three datasets and therefore the gallery size for the extended dataset experiment is fixed at two gallery images per individual. The remaining images of the subject are used as probe.

Cumulative Match Characteristics (CMC) curves are computed for each experiment and the average accuracy values are presented along with standard deviations across random subsamples. The experimental protocol for all the experiments are summarized in Table 2.1. The performance of the proposed algorithm is compared with several existing algorithms namely: FPLBP [215], Pyramid Histogram of Oriented Gradients (PHOG) [8], SIFT [126], [237], and Sparse representation [217]. Besides these methods which utilize only 2D information, a comparison is also performed with 3D-PCA based algorithm [20] which computes a subspace based on depth and grayscale information.

2.3.2 Results and Analysis: Experiment 1

Component-wise analysis: As discussed in Section II, the proposed RISE algorithm has various components: *entropy*, *saliency*, and *depth* information. The experiments are performed to analyze the effect and relevance of each component. The performance of the proposed algorithm is computed in the following six cases:

- **Case (a)** RGB-D and saliency without entropy: RGB image and depth map are used directly instead of entropy maps, i.e., $F = [F_1, F_2, F_3, F_4, F_5]$, where $F_i = D(P_i)$ instead of $F_i = D(H(P_i))$, $\forall i \in [1, 4]$.

- **Case (b)** RGB only: Only the RGB image is used to extract entropy and saliency features, i.e., $F = [F_1, F_2, F_5]$

- **Case (c)** RGB-D only: Only the entropy maps are used, saliency is not used, i.e., $F = [F_1, F_2, F_3, F_4]$

- **Case (d)** RGB and saliency without entropy: RGB information is used directly instead of entropy maps, i.e., $F = [F_1, F_2, F_5]$, where $F_i = D(P_i)$ instead of $F_i = D(H(P_i))$, $\forall i \in [1, 2]$.

- **Case (e)** RGB-D only without entropy: RGB-D information is used directly instead of entropy maps, i.e., $F = [F_1, F_2, F_3, F_4]$, where $F_i = D(P_i)$ instead of $F_i = D(H(P_i))$, $\forall i \in [1, 4]$.

- **Case (f)** RGB only without saliency: $F = [F_1, F_2]$

These cases analyze the effect of different components of the proposed algorithm on the overall performance. For example, if the descriptor performs poorly in case (a), it suggests that not using entropy maps for feature extraction is detrimental to the descriptor. Similar inferences can potentially be drawn from the results of other five cases. Comparing the performance of the proposed descriptor with entropy, saliency and depth information can also determine whether the proposed combination of components improves the face recognition performance with respect to the individual components.

The results of individual experiments are shown in Figure 2-7. It is observed that removing any of the components significantly reduces the performance of the proposed algorithm. For example, the CMC curve corresponding to case (c) shows that the contribution of including visual saliency map as an added feature is important. It is observed that saliency is relevant towards stabilizing the feature descriptor and preserving intra-class similarities. Further, in cases (d) and (e), it is observed that including depth without computing entropy performs worse than not including depth information but using entropy maps to characterize the RGB image. Intuitively, this indicates that directly using depth map results in more performance loss than not using depth at all. This is probably due to the fact that depth data from Kinect is noisy and increases intra-class variability in raw form. Overall, the proposed algorithm yields 95% rank 5 accuracy on IIIT-D database. Further, Table 2.2 shows the comparison of the proposed algorithm with existing algorithms. The results indicate that, on the IIIT-D database, the proposed algorithm is about 8% better than the second best algorithm (in this case, Sparse representation [217]). Compared with 3D-PCA algorithm, the proposed algorithm is able to yield about 12% improvement.

Fusion of algorithms: Experiments are performed with various combinations of the proposed RISE and ADM approaches as well as 3D-PCA [20]. In order to fuse 3D-PCA with RISE and ADM, both weighted sum rule and weighted Borda count can be utilized. The results of this experiment are presented in Figure 2-8. W.B.C. refers to rank level fusion using Weighted Borda Count and W.S. refers to match score level fusion using Weighted Sum rule. The key analysis are explained below:

- The proposed RISE + ADM with weighted sum rule yields the best rank 5 identification accuracy of 95.3%. RISE+ADM approach using weighted borda count also performs well providing an accuracy of 79.7% which is better than the remaining combinations at rank 1.

- Even though RISE+ADM+3D-PCA performs second best with rank 5 identification accuracy of 93.7%, the difference in performance at rank 1 is 10.9% lower than RISE+ADM (W.S.) and the use of 3D-PCA also adds to the computational complexity.

- The weighted sum variants of the combinations perform consistently better than their weighted borda count variants. This indicates that match score level fusion performs better than rank

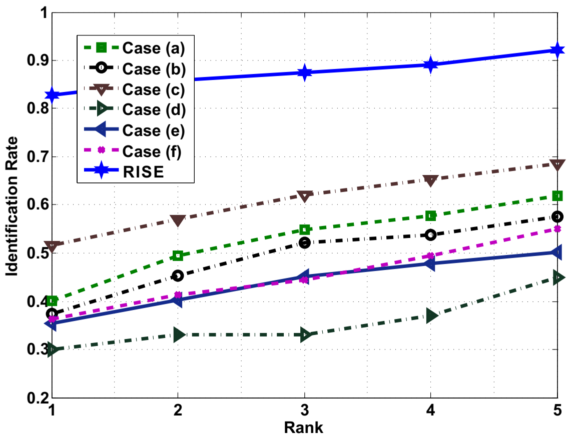


Figure 2-7: Analyzing the proposed RISE algorithm and its individual components on the IIIT-D RGB-D face database.

level fusion. However, it is also notable that the difference in performance for all approaches reduces at rank 5 compared to rank 1. This implies that any advantage gained by utilizing one approach over the other diminishes at higher ranks as the criteria for successful identification is relaxed.

Since weights are involved in both weighted borda count and weighted sum approaches, it is interesting to observe how the performance of the proposed algorithm varies with the variation in weights. The results of this experiment are presented in Figs. 2-9 and 2-10 for weighted sum rule and weighted borda count respectively. The number in parenthesis after the algorithm indicates their weight in the approach. For example, RISE (0.5) + ADM (0.5) implies that both RISE and ADM are allocated equal weights. Based on these results, the following analysis can be performed:

- The best performance is achieved with RISE (0.7) + ADM (0.3) for both the fusion algorithms. This indicates that texture features extracted by RISE are more informative for identification and therefore must be assigned higher weight. However, the geometric features from ADM also contribute towards the identification performance after fusion, thereby increasing the rank 5 accuracy from 92.2% (RISE only) to 95.3% (RISE + ADM) when weighted sum rule is utilized.

- The performance of weighted borda count is lower than weighted sum possibly because of the loss of information that occurs in using the ranked list for fusion instead of the match scores.

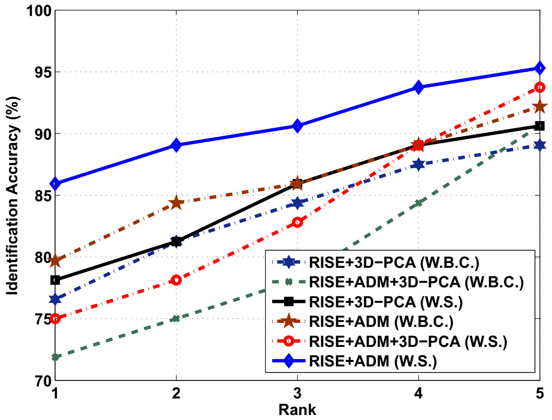


Figure 2-8: Analyzing the performance of different combinations of the proposed algorithm with 3D PCA and fusion algorithms on the IIIT-D RGB-D face database.

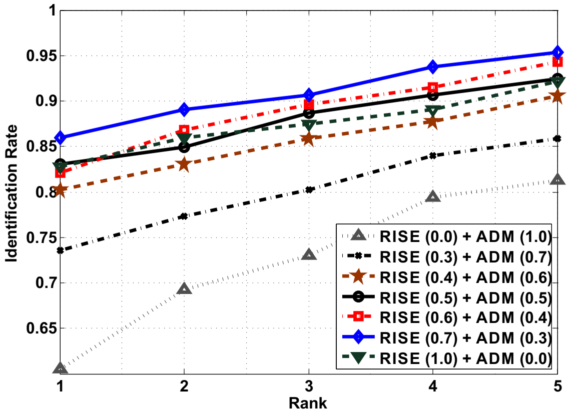


Figure 2-9: Analyzing the effect of weights in match score level fusion using weighted sum rule on the IIIT-D RGB-D face database.

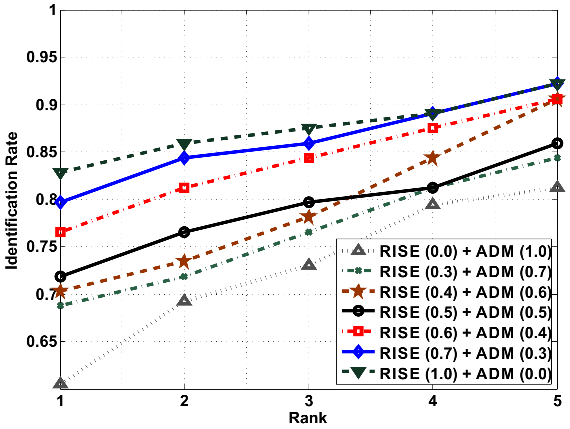


Figure 2-10: Analyzing the effect of weights in rank level fusion using weighted borda count on the IIIT-D RGB-D face database.

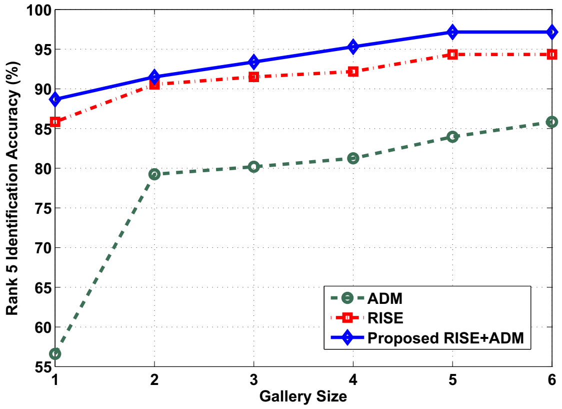


Figure 2-11: Analyzing the effect of gallery size on the identification performance on the IIIT-D RGB-D face database.

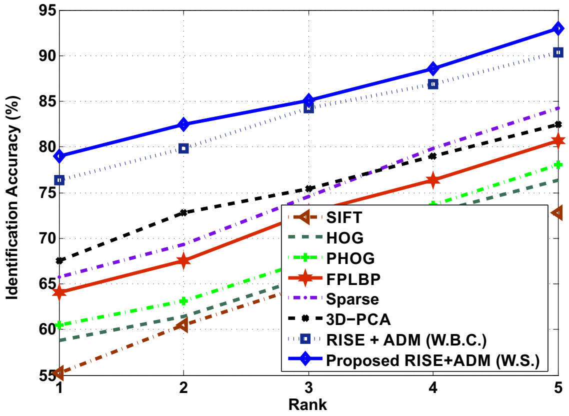


Figure 2-12: Comparing the performance of the proposed approach with existing 2D and 3D approaches on the extended database.

- Experiments have been conducted to assess the performance with all other combinations of weights as well, but none of these combinations perform better than RISE (0.7) + ADM (0.3).

Analysis with gallery size: All the experiments described above on the IIIT-D RGB-D database are performed with a gallery size of four. To analyze the effect of gallery size on the identification performance, additional experiments are performed by varying the number of images in the gallery. The results of this experiment are presented in Figure 2-11 and the analysis is presented below.

- The curve indicates that the performance of RISE, ADM and the proposed RISE+ADM approach increases with increase in gallery size. However, the maximum increment is observed from gallery size 1 to gallery size 2 in the ADM approach. This major performance increment of 22.6% can be credited to the possibility that using only single gallery image yields approximate geometric attributes. As soon as more than one sample becomes available, the averaging process increases the reliability of the geometric attributes and hence there is a significant increase in performance.

- With the above discussed exception, the performance of each approach increases consistently but in small amounts with increase in gallery size. Therefore, after a certain point, increasing gallery size does not provide high returns in terms of the performance. It is notable that even with single gallery image, the proposed algorithm yields the rank 5 identification accuracy of 89%.

Assessing the accuracy of ADM keypoint labeling: The performance of ADM approach is dependent on the keypoint labeling phase. In order to determine the accuracy of this phase, manual keypoint labels are collected via crowd-sourcing. Human volunteers are requested to label the keypoints (nose, left eye, right eye and chin) on 10 images of every subject. The average of human-annotated keypoint co-ordinates is computed and compared with the automatically obtained keypoints. An automatic keypoint is considered to be correct if it lies within a small local neighborhood of the average human-annotated keypoint. It is observed that the overall accuracy of automated keypoint labeling, using manual annotations as ground truth, on the IIIT-D Kinect

Table 2.2: Identification accuracies (%) on the IIIT-D RGB-D face database and EURECOM database individually. The mean accuracy values are reported along with standard deviation.

Modality	Descriptor	Rank 5 Identification Accuracy (%)	
		IIIT-D RGB-D	EURECOM
2D	SIFT	50.1±1.4	83.8±2.1
	HOG	75.1±0.7	89.5±0.8
	PHOG	81.6±1.4	90.5±1.0
	FPLBP	85.0±0.7	94.3±1.4
	Sparse	87.2±1.9	84.8±1.7
3D	3D-PCA	83.4±2.1	94.1±2.7
	RISE + ADM	**95.3±1.7**	**98.5±1.6**

RGB-D database is 90.1% with a 5 × 5 neighborhood and 93.6% with a neighborhood size of 7 × 7. Based on the performance of ADM on individual frames, it can be noted that it performs the best on near frontal frames and semi-frontal frames.

Performance on EURECOM: Performance of the proposed algorithm is also compared with existing algorithms on the EURECOM dataset. In order to perform this recognition experiment, the gallery sizes for the EURECOM dataset is fixed at 2 images per subject. The results of this experiment are presented in Table 2.2. The analysis is similar to the IIIT-D database and the proposed algorithm yields an accuracy of 98.5% rank-5 identification accuracy which is around 4% better than existing algorithms. Note that the EURECOM database is relatively smaller than IIIT-D database and therefore, near perfect rank 5 accuracy is achieved.

2.3.3 Results and Analysis: Experiment 2

The proposed RISE + ADM approach is compared with some existing 2D and 3D approaches on the extended dataset (Experiment 2). The identification performance of these approaches is presented in Figure 2-12 and summarized in Table 2.3. The results indicate that the proposed RISE+ADM algorithm (both weighted sum and weighted borda count versions) outperforms the existing approaches by a difference of around 8% in terms of the rank 5 identification performance. The proposed algorithm yields the best results at rank 1 with an accuracy of 78.9% which is at least 11.4% better than second best algorithm, 3D-PCA.

Detailed comparison with other algorithms: In order to compare the performance of the pro-

Table 2.3: Identification accuracy (%) for the extended gallery experiments. The mean accuracy values are reported along with standard deviation.

Modality	**Descriptor**	**Rank 1**	**Rank 5**
2D	SIFT	55.3 ± 1.7	72.8 ± 2.1
	HOG	58.8 ± 1.4	76.3 ± 1.8
	PHOG	60.5 ± 1.6	78.1 ± 1.1
	FPLBP	64.0 ± 1.1	80.7 ± 2.0
	Sparse	65.8 ± 0.6	84.2 ± 0.8
3D	3D-PCA	67.5 ± 1.2	82.5 ± 1.9
	RISE+ADM (W.B.C.)	76.3 ± 1.0	90.3 ± 1.1
	RISE+ADM (W.S.)	**78.9 ± 1.7**	**92.9 ± 1.3**

Table 2.4: A detailed comparative analysis of the proposed algorithm with 3D-PCA, FPLBP, and Sparse approaches. T and F represent True and False respectively. True ground truth refers to genuine cases and false ground truth refers to the impostor cases.

Algorithm Results	**Ground Truth**	
	True	**False**
3D-PCA=T, Proposed=T	61.9%	5.3%
3D-PCA=F, Proposed=T	21.3%	5.4%
3D-PCA=T, Proposed=F	10.0%	24.3%
3D-PCA=F, Proposed=F	6.8%	65.0%
FPLBP=T, Proposed=T	61.8%	6.8%
FPLBP=F, Proposed=T	27.6%	3.4%
FPLBP=T, Proposed=F	6.3%	25.3%
FPLBP=F, Proposed=F	4.3%	64.5%
Sparse=T, Proposed=T	68.6%	3.2%
Sparse=F, Proposed=T	18.7%	11.4%
Sparse=T, Proposed=F	8.0%	26.0%
Sparse=F, Proposed=F	4.7%	59.4%

posed algorithm with other top performing algorithms, a comparative study is performed. The details of this study are presented in Table 2.4. As is evident from the results presented, the proposed algorithm is able to correctly determine ground truth in the case of a wrong decision by another algorithm more often than the reverse case, i.e., when another algorithm is correct and the proposed algorithm is incorrect. For example, the percentage of impostor cases when 3D-PCA is incorrect and the proposed algorithm is correct is 24.30% whereas the percentage of impostor cases where the proposed algorithm is incorrect and 3D-PCA is correct is only 5.38%.

In order to further analyze the performance, we examine two types of results. Figure 2-13 contains two samples of gallery and probe images. Case 1 is when all the algorithms could match the probe to the gallery image and successfully identify the subject. Case 2 is when only the proposed algorithm is able to identify the subject and other algorithms provide incorrect results. As it can be seen from the example images of Case 1, when there are minor variations in expression and pose, all the algorithms are able to correctly identify. However, as shown in case 2, the proposed algorithm is able to recognize even when there are high pose and expression variations. Thus, it can be concluded that the proposed descriptor outperforms these existing 2D and 3D approaches. In summary, this difference in performance can be attributed to the following reasons:

- The RISE descriptor uses depth information in addition to traditional color information which allows it to utilize additional sources for feature extraction. After characterization by local entropy, the depth map is able to mitigate the effect of illumination and expression. The geometrical attributes obtained from the ADM approach further contribute towards resilient identification.

- The proposed descriptor utilizes saliency map for feature extraction to model visual attention. The saliency distribution of a face is not significantly affected by pose variations and therefore it provides tolerance to minor pose variations.

- Compared to existing approaches, entropy and saliency maps of RGB-D images are not highly affected by noise such as holes in depth map and low resolution, and therefore, yield higher performance. The additional geometric attributes are another source of noise tolerant features as they are averaged across multiple gallery images.

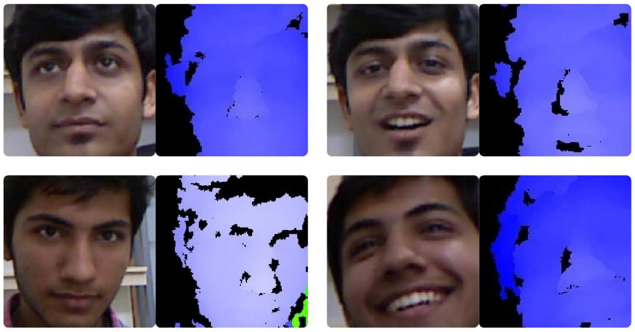


Figure 2-13: Analyzing the performance of the proposed algorithm. The first row (Case 1) presents sample gallery and probe images when all the algorithms are able to recognize. The second row (Case 2) presents example gallery and probe images when only the proposed algorithm is able to correctly identify the subject at rank-1.

Table 2.5: Rank-1 identification accuracies on the 3D TEC [206] dataset. The results of other algorithms are presented as reported in [206].

Algorithm	**Rank 1 Identification Accuracy**			
	I	**II**	**III**	**IV**
Alg. 1 (E_{pkn})	93.5%	93.0%	72.0%	72.4%
Alg. 1 (E_{minmax})	94.4%	93.5%	72.4%	72.9%
Alg. 2 (SI)	92.1%	93.0%	83.2%	83.2%
Alg. 2 (eLBP)	91.1%	93.5%	77.1%	78.5%
Alg. 2 (Range PFI)	91.6%	93.9%	68.7%	71.0%
Alg. 2 (Text, PFI)	95.8%	96.3%	91.6%	92.1%
Alg. 3	62.6%	63.6%	54.2%	59.4%
Alg. 4	98.1%	98.1%	91.6%	93.5%
Proposed	95.8%	94.3%	90.1%	92.4%

2.3.4 Experiments on 3D TEC dataset

In order to evaluate the performance of the proposed RGB-D recognition algorithm on other 3D databases, identification results are also presented on the 3D-Twins Expression Challenge (3D-TEC) dataset [206]. The database contains images pertaining to 107 pairs of twins acquired using a Minolta VIVID 910 3D scanner in controlled illumination and background. The range and texture images are of 480 × 640 resolution. The dataset provides four sets for performing identification experiments between two twins, A and B. Each set defines the gallery and probe images for each twin according to the expression variations (smile or neutral). Further details of these sets are provided in [206].

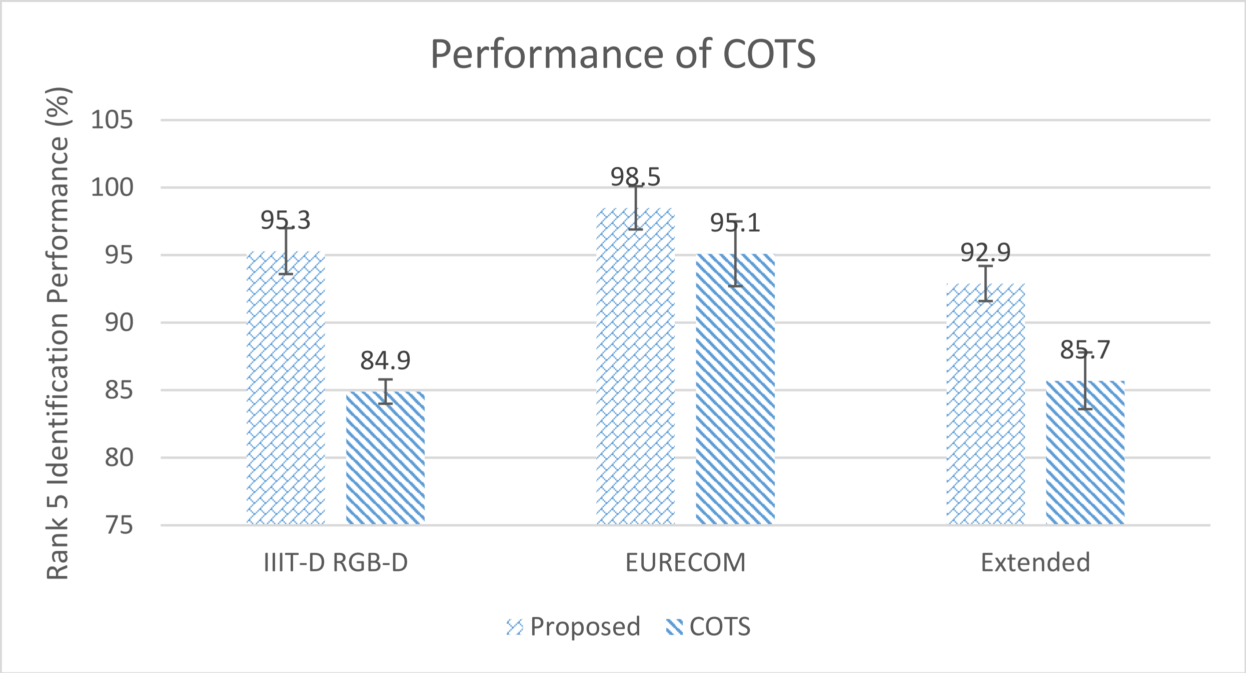


Figure 2-14: Comparing the identification performance of the proposed algorithm with COTS on all three databases.

Along with the proposed algorithm, we also compare the results with four existing algorithms that participated in the Twin Expression Challenge, 2012. The existing algorithms, Alg. 1 to Alg. 4 (named as such in the original paper), are designed to utilize rich 3D maps and/or texture information captured using telephoto lens equipped Minolta scanner. The details of these algorithms can also be found in [206]. Table 2.5 presents the results of the proposed and four existing algorithms on the 3D-TEC dataset. It is to be noted that the results of existing algorithms are taken from [206] directly. As shown in Table 2.5, even though the proposed algorithm does not fully utilize rich depth maps, it achieves the second best performance on two of the four sets and is able to yield close to state-of-the-art performance with more than 90% rank 1 accuracy on all four sets.

2.4 Summary

Existing face recognition algorithms generally utilize 2D or 3D information for recognition. However, the performance and applicability of existing face recognition algorithms is bound by the information content or cost implications. This research proposes a novel RISE algorithm that utilizes the depth information along with RGB images obtained from Kinect to improve the recognition performance. The proposed algorithm uses a combination of entropy, visual saliency, and depth information with HOG for feature extraction and random decision forest for classification. Further, the ADM algorithm is proposed to extract and match geometric attributes. ADM is then combined with the RISE algorithm for identification. The experiments performed on the RGB-D databases demonstrate the effectiveness of the proposed algorithm and show that it performs better than some existing 2D and 3D approaches of face recognition.

Appendix

The performance of the proposed algorithm is also compared with a commercial system (3D-Commerical Off-The-Shelf (COTS))[2]. COTS employs a 3D model reconstruction for each subject using the gallery RGB images. RGB probe image is also converted to 3D model for matching. The details of reconstruction algorithm are not available. Figure 2-14 presents a comparison of identification performance between COTS and the proposed algorithm. It is evident that the proposed algorithm is able to consistently achieve better performance. The failure of COTS can be attributed to the 3D reconstruction method which possibly suffers from low spatial resolution of RGB images.

[2] Name of the commercial system is suppressed due to the constraints in the license agreement.

Chapter 3

Group Sparse Representation Based Classification for Multi-feature Multimodal Biometrics

Biometrics research can be broadly classified into two categories: unimodal and multimodal. Multimodal biometrics is combining information from multiple unimodal biometric sources [94]. Researchers have shown that combining information can be beneficial when the quality or information content of one of the information sources is not sufficient for recognition. Multiple biometric information sources can be combined at different levels; namely, (a) sensor-level, (b) feature-level, (c) score-level, (d) rank-level, and (e) decision-level [94]. Fusion at each level has its advantages and limitations. For example, fusion at the sensor-level can preserve most of the information from each of the modalities however, sensor-level information may not be very discriminatory in nature [189]. While feature-level fusion does not suffer from noise to the same degree as in the case of sensor-level and also preserves much more information as compared to score-level, there exist various challenges in utilizing it. First, the relationships between different features are not always known. Second, some features are variable-length whereas others are fixed-length and therefore concatenation, which is a popular method of feature fusion [94], is not applicable in a large number of cases. Third, if these features do not reside in a commensurate space it is difficult for a classifier to determine reliable decision boundaries. Therefore, relatively less research has focused on feature-level fusion.

A review of existing biometric feature fusion algorithms in biometrics literature is presented in Table 3.1. It can be observed that feature fusion has been utilized extensively in combining features from complementary biometric modalities. Concatenation followed by feature selection or reduction is a popular approach along with distance metric learning based and discriminant analysis based algorithms.

Table 3.1: A literature review of feature-level fusion in biometrics.

Authors	Algorithm	Modalities	Features
Kumar *et al.* [108]	Feature concatenation	Palmprint and hand-geometry	Standard deviations of combined directional maps of palmprints and measurements of hand lengths and widths [108]
Chang *et al.* [24]	Combining face and ear image	Face and ear	PCA [203]
Ross and Govindarajan [176]	Feature normalization, concatenation, and performance-oriented feature selection	Face and hand	PCA [203] and LDA (face) and 9-byte features [174] (hand)
Yao *et al.* [228]	Distance-based separability weighting strategy	Face and palmprint	Gabor PCA
Rattani *et al.* [172]	Adding key-point descriptor to minutiae features, concatenating, and dimensionality reduction	Face and fingerprint	SIFT (face) [19] and minutia features [95] (fingerprint)
Singh *et al.* [189]	Adaptive SVM based fused feature selection	Face	Amplitude and phase features using 2D log polar Gabor wavelet
Zhou *et al.* [243]	Concatenation followed by multiple discriminant analysis	Side face and gait	PCA
Carvalho and Rosa [37]	Fisher's criterion based feature selection	Footstep sounds	Gait frequency, spectral envelope, cepstral and mel-cepstral analysis and loudness
Matovski *et al.* [144]	Concatenation followed by feature reduction	Gait (multiple views)	Gait energy image and gait entropy image
Krishneswari and Arumugam [103]	Fusion of the low-level features (approximation images) of both modalities prior to high-level feature extraction	Palmprint and fingerprint	Discrete Cosine Transform [6]
Rathore *et al.* [171]	Feature template fusion using set union approach	Profile face and ear	SURF [9]

Lu *et al.* [132]	Multiview neighborhood repulsed metric learning	Face	LBP [198], Linear Embedding [177], and SIFT [19]
Chai *et al.* [23]	Feature concatenation and linear discriminant analysis	Face	Gabor ordinal measures [23]
Yan *et al.* [224]	Discriminative multi-metric learning	Face	LBP [198], Spatial pyramid learning [244], and SIFT [19]
Chin *et al.* [31]	Feature concatenation	Fingerprint and palmprint	Bank of 2D Gabor filters
Goswami *et al.* [66]	Feature concatenation	Face (RGB-D)	HOG [35] of depth/visual entropy and visual saliency
Odinaka *et al.* [154]	Concatenation before and after feature selection	Cardio vascular	Electrocardiogram [155] and laser Doppler vibrometry [27]
Huang *et al.* [87]	Biometric quality based piece-wise weighted concatenation	Face and ear	PCA
Huang *et al.* [85]	Feature concatenation	Face	PCA and LDA applied on top-level's wavelet sub-bands
Kumar *et al.* [242]	Feature binarization and masking	Iris	Deeply learned spatially corresponding features
Kumar *et al.* [104]	Multi-channel QSOC feature template	Iris	Quaternian Sparse Orientation Code

Multimodal biometrics can also be beneficial when the data is captured in an unconstrained environment and there are instances of missing information. While researchers have proposed several feature fusion algorithms, not all the algorithms can efficiently combine features in the presence of missing information. The performance of popularly used feature fusion algorithms such as concatenation and PCA is significantly affected due to missing information. It is our assertion that a well designed feature-level fusion algorithm which addresses the above mentioned challenges, can aid in enhancing the state-of-the-art in biometric recognition. Since the selection of the classifier for matching is also critical towards performance and there is a high chance of discrepancy in case different feature-fusion and classification methodologies are applied, it is optimal if the classifier can handle multiple features for every data point inherently. In such a manner, the requirement to have a compatible feature-level fusion and classification technique is removed and this process is integrated in the classification stage itself.

In this research, we propose a multimodal multi-feature classifier termed as the GSRC, an extension of the existing SRC [217], which handles a multimodal multiple feature representation for

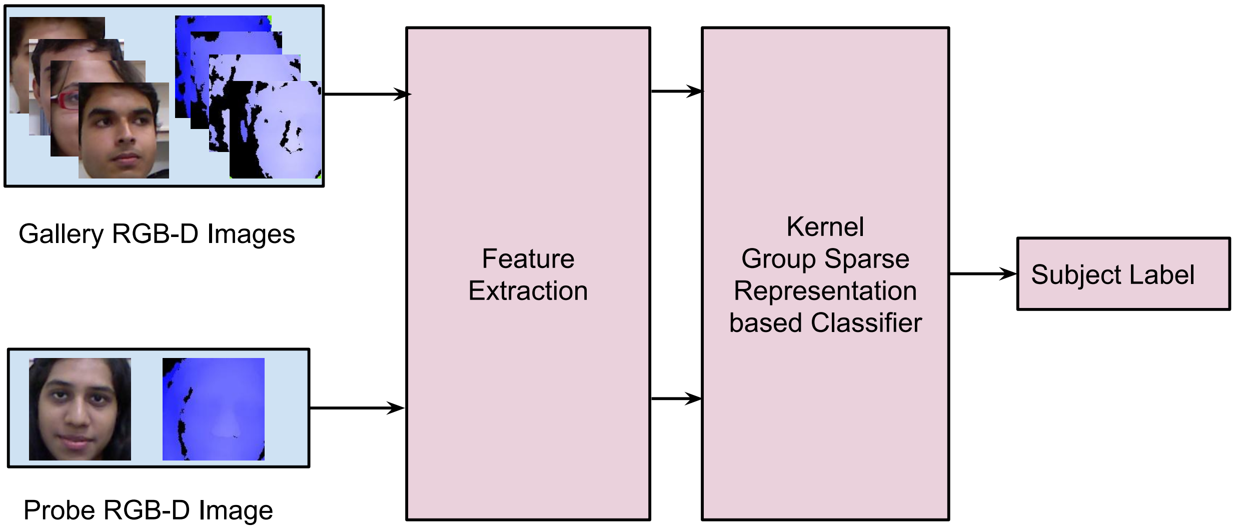


Figure 3-1: A concept diagram of the proposed algorithm.

every data point and determines the class of test data by solving a group sparsity criterion. Figure 3-1 presents the outline of the proposed algorithm for a multimodal biometric recognition scenario. Face, iris, and fingerprint modalities of a person are encoded with multiple feature representations and matched using the the proposed algorithm. By considering each feature source without the use of concatenation or feature reduction, the classification algorithm can utilize the different feature spaces to make an optimal decision. While a few extensions to the traditional SRC have been proposed in literature [86, 104, 105, 186], the proposed algorithm presents an alternate perspective towards group sparsity. The performance of the proposed GSRC classifier is evaluated with multiple feature sets and biometric modalities on two databases: publicly available WVU multimodal biometric dataset [34] and a real world database obtained from a Law Enforcement Agency [15]. The performance is also compared with existing state-of-the-art algorithms.

The group sparse classifier represents each test sample as a combination of its representations in individual feature spaces and classifies based on the residual error for each class. The quality of classification depends on the intra-class stability and inter-class differentiability of these individual representations. Projection into a higher dimensional space using an appropriate kernel function is a popular technique to achieve better separability of data. If each feature space is projected using a suitable kernel, the quality of the combined representation should improve. It is our assertion that the accuracy of the group sparse classifier can be further improved using kernelization without decreasing computational efficiency. Even though Gao *et al.* [56] have discussed kernelization in a sparse representation based approach with good results, they have not considered a multi-feature or multi-modal scenario in which case the algorithm would have to rely on score level or decision level fusion rules unlike the proposed algorithm. In this research, we also explore the applicability of kernelization to the group sparse representation based classifier. We evaluate the proposed algorithm on a variety of biometrics problems such as recognition in surveillance images in a cross-distance matching scenario, face recognition for RGB-D images obtained using Kinect sensor (as shown in Figure 1), and multi-modal biometrics using face, iris, and fingerprint.

3.1 Preliminaries

In this subsection, we briefly discuss the basic concepts of sparse representation and some recent extension of sparse representation for joint representation and non-linear representation.

3.1.1 Sparse Representation based Classification

Sparse representation based classification [217] assumes that the training samples of a particular class approximately form a linear basis for a new test sample belonging to the same class. Let v_{test} be the test sample belonging to the k^{th} class, it can be represented as,

$$v_{test} = \alpha_{k,1} v_{k,1} + \alpha_{k,2} v_{k,2} + \cdots + \alpha_{k,n} v_{k,n} + \epsilon \tag{3.1}$$

where, $v_{k,i}$ denotes the i^{th} training sample and ϵ is the approximation error.

In a classification problem, the training samples and their class labels are provided. The task is to assign the given test sample with the correct class label. This requires finding the coefficients $\alpha_{k,i}$ in Equation 3.1. Since the correct class is not known, SRC represents the test sample as a linear combination of all training samples from all classes,

$$v_{test} = V\alpha + \epsilon \tag{3.2}$$

where, $V = \left[\underbrace{v_{1,1}|\dots|v_{1,n}}_{v_1} | \underbrace{v_{2,1}|\dots|v_{2,n}}_{v_2} | \dots \underbrace{v_{c,1}|\dots|v_{c,n}}_{v_c} \right]$ and

$\alpha = \left[\underbrace{\alpha_{1,1},\dots,\alpha_{1,n}}_{\alpha_1}, \underbrace{\alpha_{2,1},\dots,\alpha_{2,n}}_{\alpha_2}, \dots \underbrace{\alpha_{c,1},\dots|\alpha_{c,n}}_{\alpha_c} \right]$

According to SRC, only the training samples from the correct class should form the basis for representing the test sample and the samples from other classes should not contribute. Based on this assumption, it is likely that vector α is sparse, *i.e.*, it should have non-zero values corresponding to the correct class and zero values for other classes. Thus Equation 3.2 is a linear inverse problem with a sparse solution. In [217], the coefficient α is solved by employing a sparsity promoting l_1-norm minimization.

$$\min_{\alpha} ||v_{test} - V\alpha||_2^2 + \lambda||\alpha||_1 \tag{3.3}$$

$||\alpha||_1$ denotes the l_1 norm of α which is $\sum_i^N |\alpha_i|$, where $|\cdot|$ denotes the absolute value function and N represents the length of vector α. With the sparse solution of Equation 3.3, Wright *et al.* [217] proposed the following algorithm to determine the class of the test sample.

1. Solve the optimization problem in Equation 3.3.

2. For each class k repeat the following two steps:

   - Reconstruct a sample for each class by a linear combination of training samples belonging to that class by the equation $v_{recon}(k) = V_k\alpha_k$.

   - Find the error between the reconstructed sample and the given test sample by $error(v_{test}, k) = ||v_{test} - v_{recon}(k)||_2$.

3. Once the error for every class is obtained, assign the test sample to the class having the minimum error.

The assumption here is that the representative sample for the correct class will be similar to the test sample. However, there can be several other ways to assign the class. Li and Lu showed that considering the magnitude of the coefficient α can yield better results [119]. Assuming that the SRC assumption holds true, values in α corresponding to the correct class should only be true,

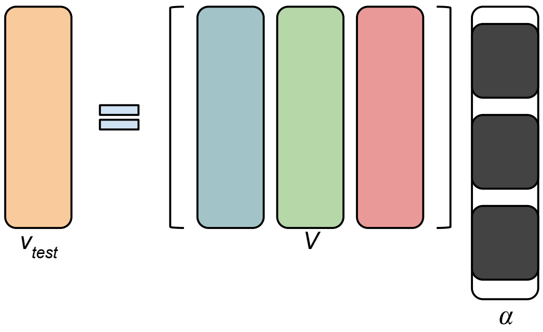


Figure 3-2: Illustrating the Sparse Inverse Problem

and the values for other classes should be zero or close to zero. In [119], $||\alpha_k||_1$ for each class was computed and the test sample was assigned to the class which had the highest value.

The diagrammatic representation of the inverse problem with a sparse solution is shown in Figure 3-2. The forward problem (assumption) is such that the vector v_{test} is formed by a linear combination of a few columns (training samples of correct class) of V. In the inverse problem, given the test sample v_{test}, few columns of V that are used to represent the test sample are selected and the corresponding sparse coefficient α is found.

In Figure 3-2, the colored blocks in V represent a subspace. Therefore, v_{test} can be modeled as a union of subspaces. Analyzing l_1-minimization as finding the union of subspaces is a powerful tool for studying the success of SRC. The SRC has been successfully applied for face recognition [217]. In a simplistic view, one can assume that for each personÃć a face can be represented roughly by three subspacesÃć frontal, left profile and right profile. A test (face) sample falls in either one of these subspaces. The l_1-norm minimization basically selects the corresponding subspace and the corresponding coefficients. For the correct subspace, the residual error (ϵ) is small. Therefore, assigning the test sample based on a small residual error is a sound criterion.

3.1.2 Block/Joint Sparse Classification

The SRC employs an l_1-minimization for solving the inverse problem. This is an unsupervised approach and it does not utilize information about the class labels. In [47, 137, 233], it is argued that α is supposed to be non-zero for all training samples corresponding to the correct class. The SRC assumes that the training samples for the correct class will be automatically selected by imposing the sparsity inducing l_1-norm; it does not explicitly impose the constraint that if one class is selected, all the training samples corresponding to that class should have corresponding non-zero values in α. [47, 137, 233] claim that it can be better recovered if the selection of all the training samples within the class is enforced. This is achieved by employing a supervised $l_{2,1}$-norm instead of the l_1-norm.

$$\min_{\alpha} ||v_{test} - V\alpha|| + \lambda||\alpha||_{2,1} \tag{3.4}$$

Here, the mixed norm is defined as:

$$||\alpha||_{2,1} = \sum_{k=1}^{n} ||\alpha_k||_2 \tag{3.5}$$

The inner l_2-norm enforces the selection of all the training samples within the class, but the sum of l_2-norm over the classes acts as l_1-norm over the selection of classes and selects very few classes. The block sparsity promoting $l_{2,1}$-norm ensures that if a class is selected, all the training samples within the class are used to represent the test sample.

The Block Sparse representation-based Classification (BSC) approach is effective for general purpose classification problems and is shown to perform well for simple classification problems [47, 137, 233]. However, it yields very low accuracies compared to SRC for face recognition. To analyze this phenomenon we refer to Figure 3-2, in BSC all the training samples from the same class have the same class label. Therefore, the $l_{2,1}$-minimization attempts to select all the training samples to represent the test sample. It considers all the colored blocks in Figure 3-2 as a single subspace instead of a union of subspaces; which may not be correct approach in all the situations. Enforcing block sparsity is a good idea when the classification problem is simple and all the samples truly belong to a single subspace, e.g. in fingerprint recognition or character recognition. It prevents selection of samples from arbitrary classes. However, face recognition does not satisfy this simplistic assumption. As mentioned before, the face images can belong to three subspaces. The BSC tries to combine all the subspaces into a single one, for instance, if the test sample is a left profile, it will try to fit the left and right profiles as well as the frontal view to the test sample. This is clearly an error prone technique and it has been observed that BSC fails for face recognition related problems, especially in challenging situations with a large variability in the training and test samples.

3.1.3 Non-Linear Extensions

In [136, 139], non-linear extensions to the SRC [139] and BSC [136] are proposed. The linearity assumption is generalized to include non-linear (polynomial) combinations. The generalization of Equation 3.2 leads to:

$$v_{test} = f(V\alpha) + \epsilon \tag{3.6}$$

Here, f denotes a non-linear function and ϵ denotes the approximation error. The assumption is that the test sample can be represented as a non-linear combination of the training samples. Notice that this is different from the kernel-based techniques. In these studies, the recovery of the coefficient vector requires solving a non-linear inverse problem with sparsity constraints,

$$\min_{\alpha} ||v_{test} - f(V\alpha)||_2^2 + \lambda||\alpha||_1 \tag{3.7}$$

There are no off-the-shelf solutions to solve Equation 3.7. In [136, 139], FOCally Underdetermined System Solver (FOCUSS) and Orthogonal Matching Pursuit (OMP) based solvers are modified to accommodate the non-linearity. The non-linear extension shows good results on generic classification problems. Several researchers proposed the Kernel Sparse Representation based Classification (KSRC) approach [28, 230, 239]. KSRC is a simple extension of the SRC using the Kernel trick. The assumption here is that the non-linear function of the test-sample can be represented as a linear combination of the non-linear functions of the training samples, i.e.

$$\phi(v_{test}) = \phi(V)\alpha + \epsilon \tag{3.8}$$

Here, $\phi(\cdot)$ represents a non-linear function. The simplest way to apply the kernel trick is to pre-multiply by $\phi(V)^\intercal$.

$$\phi(V)^\intercal\phi(v_{test}) = \phi(V)^\intercal\phi(V)\alpha + \epsilon \tag{3.9}$$

The expression in Equation 3.8 consists of inner products between the training samples and the test sample on the left hand side and inner products between the training samples on the right hand side. Once we have the representation in terms of inner products, the kernel-trick can be applied as follows,

$$k(x_i, x_j) = \left\langle \phi(x_i), \phi(x_j) \right\rangle \tag{3.10}$$

Here, $\langle .... \rangle$ represents the inner product. Applying the kernel trick allows representing Equation 3.9 in the following form,

$$v^k_{test} = (V)^k \alpha + \epsilon \tag{3.11}$$

Here, the superscript k represents the kernelized version of the test sample and training data. Equation 3.11 can be solved using any standard l_1-solver. The elegant formulation of the kernel trick we have discussed here was proposed in [230]. In other studies, the sparsity promoting solver (l_1-minimization or OMP) was modified to accommodate the kernel trick.

3.1.4 Extension to Multiple Measurement Vectors

So far we have discussed scenarios where the objective is to classify a single instance of the test sample. This is known as the Single Measurement Vector (SMV) problem. In many scenarios the test sample naturally consists of multiple samples, e.g. videos or multi-spectral imaging. These are referred to as the Multiple Measurement Vector (MMV) problem.

Consider the problem of video based face recognition [138]. Here, the training data V consists of video frames for a person. There are multiple training samples in each class; the structure of V is therefore the same as in SMV problems. However, the test sample consists of a video sequence containing multiple frames (T frames), $V_{test} = [v^1_{test}|v^2_{test}| \dots |v^n_{test}]$. Assuming that the SRC assumption holds for each frame, i.e.

$$v^i_{test} = (V)\alpha^{(i)} + \epsilon \quad \forall i \in \{1 \dots T\} \tag{3.12}$$

It can be combined for all the frames in the following manner,

$$V_{test} = VZ + \epsilon \tag{3.13}$$

where, $Z = [\alpha_1| \dots |\alpha_T]$

According to the SRC assumption each of the $\alpha^{(i)}$'s should be sparse. The non-zero values correspond to the training samples of the correct class. Every $\alpha^{(i)}$ is supposed to be represented by training samples of the correct class. Therefore, the sparsity signature (positions of non-zero values) is expected to remain the same in every $\alpha^{(i)}$. If this assumption holds true, then Z is supposed to be row-sparse and only those rows that correspond to the correct class will have non-zero values. The row-sparsity assumption leads to solving the inverse problem in Equation 3.13 via:

$$\min_{Z} ||V_{test} - VZ||_F^2 + \lambda ||Z||_{2,1} \tag{3.14}$$

where, $||Z||_{2,1} = \sum\limits_{j} ||Z^{j\rightarrow}||_2^2$ and $Z^{j\rightarrow}$ represents the j^{th} row of Z.

The argument for using the mixed $l_{2,1}$-norm is the same as before. The l_2-norms over the rows promote a dense (non-zero) solution within the selected row but the outer l_1-norm enforces sparsity on the selection of rows. It should be noted that the $l_{2,1}$-norm in the current case is defined on a matrix and it should not be confused with the BSC assumption of block sparsity where it was defined on a vector. The final classification decision is similar to the SRC approach. Once Z is recovered, the class representation can be obtained by partitioning Z according to the classes. The MMV test sample V_{test} can be assigned to the class having the minimum residual error.

3.2 Proposed Group Sparse Representation Based Classification

The proposed GSRC algorithm is a generic classification algorithm that can handle multiple features and data sources for each data point. In this research, we propose the formulation and discuss its application for the problem of multimodal biometrics.

Let N be the number of biometric modalities; for each modality, we assume that the sparse representation classification model holds true, i.e., the test sample from that modality can be expressed as a linear combination of the training samples of the correct class from the same modality.

$$v_{test}^i = V^i \alpha^i + \epsilon \quad \forall i \in \{1 \ldots N\} \tag{3.15}$$

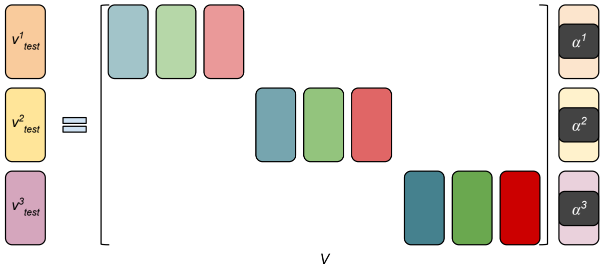


Figure 3-3: Illustrating the proposed GSRC algorithm.

It is possible to solve each of modalities using the SRC algorithm and combine them at a later stage using a score level fusion rule. However, such an approach does not exploit the intrinsic structure of the problem. A better approach is to combine all the modalities into a single framework. As shown in Figure 3-3, this can be succinctly represented as:

$$\begin{bmatrix} v_{test}^{1} \\ \ldots \\ \ldots \\ v_{test}^{N} \end{bmatrix} = \begin{bmatrix} V^{1} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots \\ 0 & \ldots & V^{N} \end{bmatrix} \begin{bmatrix} \alpha^{1} \\ \ldots \\ \ldots \\ \alpha^{N} \end{bmatrix} + \epsilon \qquad (3.16)$$

Since each of the $\alpha^{(i)}$'s are sparse, the simplest way to solve Equation 3.16 is to impose a sparsity penalty and solve it via l_1-minimization. However, such a naive approach is sub-optimal and does not exploit the underlying structure of the problem either.

The coefficient vector (represented as a row vector for simplicity) for each modality can be expanded as: $\alpha^{i} = \left[\alpha_{1}^{i}, \ldots, \alpha_{k}^{i}, \ldots, \alpha_{c}^{i}\right]$ where, α_{k}^{i} denotes the coefficients corresponding to the k^{th} class for the i^{th} modality. If a test sample belongs to the k^{th} class, the corresponding coefficients are non-zero.

$$\alpha = \Bigg[\underbrace{\alpha_{1}^{1}, \ldots, \alpha_{k}^{1}, \ldots, \alpha_{c}^{1}}_{\alpha^{1}}, \ldots, \ldots, \underbrace{\alpha_{1}^{N}, \ldots, \alpha_{k}^{N}, \ldots, \alpha_{c}^{N}}_{\alpha^{N}} \Bigg]$$

Since the SRC assumptions holds true for individual modalities, the α_{k}^{i}'s for each i^{th} (modality) have non-zero values. Therefore, α has a group sparse structure where the non-zero elements occur corresponding to the indices of the k^{th} class. This leads to a group sparse representation where the grouping is simply based on the indices. Equation 3.16 can be solved using the group sparsity promoting $l_{2,1}$-norm:

$$\min_{Z} ||v_{test} - V\alpha||_2^2 + \lambda||\alpha||_{2,1} \tag{3.17}$$

where,

$$v_{test} = \begin{bmatrix} v_{test}^1 \\ \dots \\ \dots \\ v_{test}^N \end{bmatrix}, V = \begin{bmatrix} V^1 & \dots & 0 \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ 0 & \dots & V^N \end{bmatrix} and \quad \alpha = \begin{bmatrix} \alpha^1 \\ \dots \\ \dots \\ \alpha^N \end{bmatrix} \tag{3.18}$$

The proposed GSRC formulation does not suffer from these limitations that fraught block sparse classification [137, 233]. Here, we are not trying to fit one vector (test sample) to all the subspaces simultaneously (as is done by BSC); but we are fitting test samples from each modality into the subspaces spanned by the training samples of the same modality. In simpler words, the main difference between our proposition and previous studies is that we define the group based on indices from different modalities whereas previous studies define the group based on class labels. The $l_{2,1}$-norm has also been utilized in the sparse representation literature in different ways where its equivalence to hypercomplex sparse coding is leveraged to extract multi-channel quaternionic sparse representations of iris orientation features [104]. The group sparsity constraint is applied while optimizing the dictionary coefficient vector for the individual channel encoding. In contrast, in the proposed algorithm, we apply the group sparsity constraint on the multi-modal multi-feature representation matrix and enforce group sparsity at the index level itself. Our formulation keeps the flexibility of the SRC approach and improves upon it by exploiting the multimodal biometrics problem structure. The representative sample for each class for all the modalities is computed as:

$$v_{rep}(k) = \begin{bmatrix} V_k^i \alpha_k^i \\ \dots \\ \dots \\ V_k^N \alpha_k^N \end{bmatrix} \tag{3.19}$$

The classification is based on the same principle as SRC. The test sample is assigned to the class having the minimum residual error between the test vector and the class representative. One can also use the sum of l_1-norm of the α_k^i's for each class and assign the test sample to the class having the maximum value that is in agreement with the proposal in [119]. Regardless of the criterion (minimum residual or maximum coefficient), GSRC utilizes an elegant decision rule that does not require score level fusion strategies.

3.2.1 Multi-feature Classification

The above discussion pertaining to multimodal biometrics also pertains to multi-feature classification problems in biometrics as well as other research areas. For instance, consider the problem of object recognition. Shape descriptors such as HOG and LBP provide complementary feature information regarding the images. In a recent paper [157], it was shown that these features can be used by independent sparse representation classifiers whose outputs can be used by heuristic decision rules to come up with the final decision (class label corresponding to the test sample). The proposed GSRC algorithm is an elegant and robust solution to multi-feature classification problem that incorporates all the available information (multiple feature sets) into the classification problem and yields a class label for the test data. Next, we describe the formulation of GRCS for multi-feature classification.

Let T denote the number of features; for each feature, the SRC assumption holds, i.e.

$$v_{test}^j = V^j \alpha^j + \epsilon \quad \forall j \in \{1 \ldots T\} \tag{3.20}$$

In Ouyang *et al.* [157], SRC is applied to the individual features and the predicted class is based on fusing the outputs from multiple sparse classifiers. Here, we follow the same approach as in the multimodal biometrics formulation. The combined multi-feature problem can be expressed as follows:

$$v = V\alpha + \epsilon \tag{3.21}$$

where,

$$v_{test} = \begin{bmatrix} v_{test}^1 \\ ... \\ ... \\ v_{test}^T \end{bmatrix}, V = \begin{bmatrix} V^1 & ... & 0 \\ ... & ... & ... \\ ... & ... & ... \\ 0 & ... & V^T \end{bmatrix} \, and \quad \alpha = \begin{bmatrix} \alpha^1 \\ ... \\ ... \\ \alpha^T \end{bmatrix}$$

Similar to the previous discussion, the coefficient vector α is group sparse. Assuming that k is the correct class, the test sample from each feature set is represented by the training samples from the k^{th} class of the same feature set. Therefore the k^{th} class is active (non zero) for all the feature sets. Hence, when grouped according to the indices, the coefficient vector α is group sparse. Thus, Equation 3.21 can be solved using the $l_{2,1}$-norm minimization. The subsequent decision (class prediction) follows in exactly the same manner as for multimodal biometrics.

3.2.2 Combined Multimodal and Multi-feature Classification

In the most general formulation, the multimodal and multi-feature classification problems can be combined within the GSRC framework. We assume that there are N modalities; the index for the modalities being i. For each modality, we assume that there are T_i feature sets. The SRC assumption holds for each feature set in every modality, i.e.,

$$v_{test}^{i,j} = V^{i,j}\alpha^{i,j} + \epsilon \quad \forall j \in \{1 \dots T_i\} \quad and \quad \forall i \in \{1 \dots N\} \tag{3.22}$$

As before, each feature set can be individually solved using the SRC. However, this approach does not account for the structure of the problem. Therefore, it is best to combine all the information into a single problem as follows:

$$\begin{bmatrix} v_{test}^1 \\ ... \\ ... \\ v_{test}^N \end{bmatrix} = \begin{bmatrix} V^1 & ... & 0 \\ ... & ... & ... \\ ... & ... & ... \\ 0 & ... & V^N \end{bmatrix} \begin{bmatrix} \alpha^1 \\ ... \\ ... \\ \alpha^N \end{bmatrix} + \epsilon \tag{3.23}$$

$$\text{where, } v^i_{test} = \begin{bmatrix} v^{i,1}_{test} \\ \cdots \\ \cdots \\ v^{i,T_i}_{test} \end{bmatrix}, V = \begin{bmatrix} V^{i,1} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots \\ 0 & \ldots & V^{i,T_i} \end{bmatrix} \; and \quad \alpha = \begin{bmatrix} \alpha^{i,1} \\ \cdots \\ \cdots \\ \alpha^{i,T_i} \end{bmatrix}$$

The argument presented in the previous subsections also holds here. For each feature in every modality, only a few coefficients (corresponding to the k^{th} class) are expected to be non-zero. Therefore, in every combination of modality and feature set, only the coefficients corresponding to the k^{th} class (assuming it is the correct class) are active. The full coefficient vector is group sparse when grouped by indices that are non-zero only for indices corresponding to the k^{th} class. This would allow solving Equation 3.23 by $l_{2,1}$-minimization.

Using the proposed algorithm, the process for class prediction is the same as before. The coefficient corresponding to every class is extracted to form the representative sample. The residual error between the class representative and the test vector is computed and the test vector is assigned to the class having the lowest residual.

We should reiterate that GSRC does not try to represent the test sample by all the subspaces from all modality feature set combinations, it only fits the few subspaces (corresponding to the correct class) to its corresponding feature set for a given modality. Thus, it does not violate or constrain the SRC assumption. GSRC is an elegant extension of the SRC technique that can handle multiple modalities and multiple feature sets in a single framework. Rather the SRC is a special case of the GSRC when there is only one modality and one feature set. As discussed previously, the formulation of the proposed algorithm allows us to handle missing features, which may not be the case with previous approaches such as [36] where a joint sparse MMV recovery approach is utilized.

3.3 Proposed KGSRC Algorithm

In the formulation of the GSRC algorithm, the representative sample for each subject for all the sources is computed as:

$$I_{rep}(k) = \begin{bmatrix} G_k^{s1}\alpha_k^{s1} \\ \\ G_k^{s2}\alpha_k^{s2} \end{bmatrix} \tag{3.24}$$

The test sample is assigned to the subject having the minimum residual error between the test vector and the subject representative. Multimodal biometrics features might not be linearly separable in the feature space itself. This would lead to decreased performance due to the unavailability of an ideal decision boundary to separate samples belonging to the individual subjects. However, by utilizing kernelization, these features are projected to a higher dimensional space where they may be linearly separable, thereby resulting in better decision boundaries and better performance. For each source, we can have the equivalent non-linear representation:

$$\phi(I_{probe}^s) = \phi(G^s)\alpha^s + \epsilon, \quad \forall s \in \{s1, s2\} \tag{3.25}$$

Equation 3.25 can be kernelized by pre-multiplication of $\phi(I_{probe}^s)$

$$\phi(G^s)^T\phi(I_{probe}^s) = \phi(G^s)^T\phi(G^s)\alpha^s + \epsilon, \;\; \forall s \in \{s1, s2\} \tag{3.26}$$

The kernel κ is defined as:

$$\kappa(x_i, x_j) = \Big\langle \phi(x_i), \phi(x_j) \Big\rangle \tag{3.27}$$

where $\Big\langle .... \Big\rangle$ represents the inner product. In subsequent equations, we use the notation $\kappa(x)$ to denote $\kappa(G^s, x)$. Using Equation 3.27, Equation 3.26 can be written as:

$$\kappa(I_{probe}^s) = \kappa(G^s)\alpha^s + \epsilon \tag{3.28}$$

As in GSRC, Equation 3.28 can be compactly represented as,

$$\kappa(I_{probe}) = \kappa(G)\alpha + \epsilon \tag{3.29}$$

where, $\kappa(I_{probe}) = \begin{bmatrix} \kappa\left(I^{s1}_{probe}\right) \\ \\ \kappa\left(I^{s2}_{probe}\right) \end{bmatrix}$ $and\ \kappa(G) = \begin{bmatrix} \kappa(G^{s1}) & 0 \\ & \\ 0 & \kappa(G^{s2}) \end{bmatrix}$

Similar to the GSRC approach, Equation 3.29 can be solved using $L_{2,1}$ minimization which promotes group sparsity:

$$\min_{\alpha} ||\kappa(I_{probe}) - \kappa(G)\alpha||_2^2 + \lambda||\alpha||_{2,1} \tag{3.30}$$

Once α is obtained, the representative for each subject is computed using Equation 3.24 and the probe image is assigned to the subject/class with the minimum residual. An important problem in a kernel based approach is the selection of the kernel function and its parameters since these selections can greatly influence the final performance of the algorithm. Usually, in the case of existing machine learning approaches such as SVMs that utilize kernels, the kernel and its parameters are chosen empirically. There are certain techniques such as grid search, where a subset of possible parameter values for each kernel choice can be used for training and then optimized on the basis of maximum performance on a separate validation set. However, this requires enough training data for both the training and validation sets. The validation set is kept separate from the training set since optimizing training performance alone can lead to overfitting and reduction in performance on the test set.

We propose a method to select the kernel and its parameters using just one training set that is also able to preserve robustness and avoid overfitting, thereby avoiding the requirement for a large amount of data to create a separate validation set. Using the proposed kernel/parameter selection, the KGSRC algorithm is able to remove the need for a user to fine-tune the kernel selection. The user's primary concern is then to just select the appropriate feature sources and training data, which become the parameters of the algorithm while a large set of possible kernels and associated parameters can be set as the search space for the kernel selection algorithm. To preserve robustness while performing parameter selection using the training set, we propose a metric to quantize the goodness of separation created by the transformation enacted by a kernel function. Given a list of kernel functions, K and a set of associated parameters θ_K, we compute the following metric for each possible combination for the training data:

$$\Psi_{K,\theta_{Ki}} = \frac{\Sigma_{a=1}^{C}\Sigma_{b=1}^{C}\chi^2\left(\mu_{Ki}(a), \mu_{Ki}(b)\right)}{\Sigma_{z=1}^{C}\Sigma_{x=0}^{n_z}\Sigma_{y=0}^{n_z}\chi^2(K_{\theta_i}(\gamma_{zx}), K_{\theta_i}(\gamma_{zy}))} \tag{3.31}$$

Here, θ_{Ki} denotes the i^{th} parameter setting for the kernel K, $K_{\theta_i}(\cdot)$ is the kernel function when using θ_{Ki} as the parameter setting, $\chi^2(x, y)$ denotes the χ^2 distance between x and y, C is the number of classes, n_z is the number of training samples that belong to a particular class z, γ_{zx} denotes the x^{th} training sample belonging to the z^{th} class, and $\mu_{Ki}(a)$ denotes the mean transformed vectors of class a after kernelization using the kernel K with parameter setting θ_{Ki}. The class mean for a class c is computed as follows:

$$\mu_{Ki}(c) = \frac{\Sigma_{x=0}^{x=n_c} K_{\theta_i}(\gamma_{cx})}{n_c} \tag{3.32}$$

Essentially, Equation 3.31 computes the joint kernel-parameter combination score Ψ which is a measure of how well a particular kernel-parameter function can optimize the trade-off between inter-class distance (numerator in Equation 3.31) and intra-class distance (denominator in Equation 3.31). By using the aggregated statistic for intra-class distance across all data points in the training data and using only the means to compute the inter-class distance, we help make the selection robust towards overfitting to the training data. Our assertion is that a kernel-parameter combination with a high Ψ score is one that is able to maximize this metric consistently over the entirety of the training data and should also perform well on the test data. The underlying assumption is that the test data is not drastically different from the training data which is required inherently for good performance of a sparse representation based approach even without the involvement of kernel selection. In all the discussed results, we evaluate the performance of the proposed algorithm with the proposed kernel selection approach being used to determine both the kernel and its parameters.

3.4 Experimental Results and Analysis

This section details the databases and experimental protocol, followed by experimental results and analysis. The proposed Group Sparse Representation Classifier or GSRC algorithm is evaluated on two publicly available multimodal biometric databases:

- **WVU multimodal database**: The WVU multimodal database [34] consists of data pertaining to iris, fingerprint, palmprint, hand geometry, face video and voice, and face modalities for 270 individuals. The database also includes soft biometric information such as height, weight, ethnicity, and gender. In this research, we focus on three biometric modalities: iris, face, and fingerprint. For some individuals, not all biometric samples are available, these are treated as cases of missing data and information about the concerned (missing) modality is not utilized for recognizing these probe images. Images pertaining to 108 subjects (40%) are utilized for training and data for the remaining 162 subjects (60%) is utilized for testing. Three images are used as gallery and the remaining images are used as probes. The number of images available per modality varies and therefore the number of probe images varies in the range of 680 to 6300.

- **LEA multimodal database**: The LEA database [15] contains unconstrained multimodal biometric data pertaining to 18,000 individuals. The database comprises of the face, fingerprint, and iris modalities. Similar to the WVU database, data for all three modalities is not available for each individual and hence the database encompasses all biometric covariates as well as the missing data problem. Data pertaining to 50% of the individuals, i.e., 9000, is utilized for training and the remaining 9000 individuals are utilized for testing. Two images from each individual are used as gallery and the remaining images (1-3 images per person) are used as probes.

3.4.1 Algorithms used for Performance Evaluation

In order to evaluate the performance in a multi-feature multimodality setting, two features are considered for each modality. Uniform Circular LBP (UCLBP) [156] and Speeded Up Robust Features (SURF) [9] are considered for face, Video-based Automatic System for Iris Recognition

Table 3.2: Rank-1 identification accuracy (%) with individual features and their combination (SRC and GSRC) on the WVU and LEA databases.

Modality	Features		WVU	LEA
Face	Individual	UCLBP	75.4	24.2
		SURF	79.1	28.4
	Fusion	SRC	82.3	39.7
		GSRC	83.7	40.9
Iris	Individual	Vasir	85.0	31.0
		LPG	90.5	36.4
	Fusion	SRC	92.9	41.2
		GSRC	93.5	43.5
Finger	Individual	NBIS	85.9	40.1
		VeriFinger	90.7	45.7
	Fusion	SRC	92.6	51.8
		GSRC	93.1	53.5

(VASIR) [221] and Log Polar Gabor (LPG) [205] are considered for iris, and NIST Biometric Image Software (NBIS) [1] and VeriFinger (VF)[2] are used for the fingerprint modality. We use the two-stage iris segmentation algorithm proposed in [205], in which first the inner and outer boundaries of the iris are estimated using an elliptical model. Then, the modified MumfordâĂŞShah functional [202] is applied in a narrow band over the boundaries estimated in stage one to perform exact segmentation of the iris. The performance of the proposed GSRC algorithm is compared with the SRC algorithm [217] and the state-of-the-art multimodal algorithm which is based on Context Switching [15]. Further, we utilize sum rule match score fusion [94] for performance comparison.

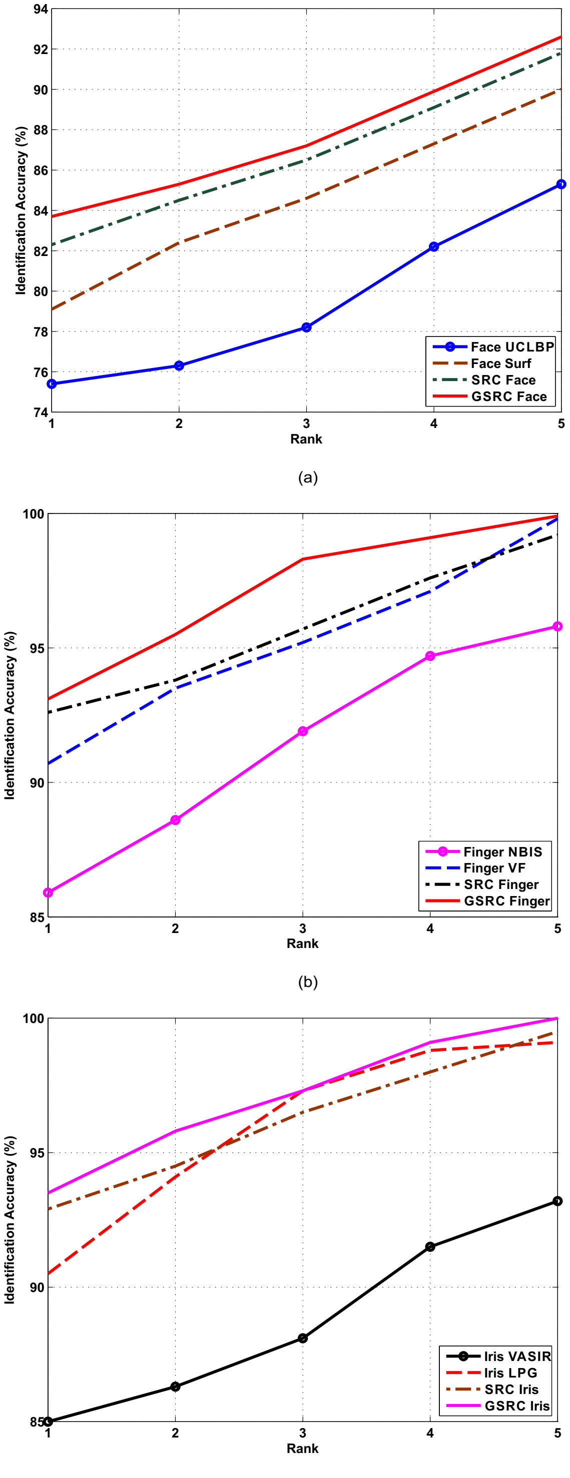
3.4.2 Results and Analysis

Identification experiments are performed on both the WVU and LEA databases and the performance of the proposed GSRC algorithm is evaluated in four scenarios and major observations are noted below. All the experimental results are presented in the form of CMC curves in Figures 3-4, 3-5, 3-6, and 3-7 and also summarized in Tables 3.2 and 3.3.

- **Single-feature single-modality**: These experiments are performed to assess the baseline performance of the individual features. As mentioned before, two features are considered

[1] http://www.nist.gov/itl/iad/ig/nbis.cfm

[2] http://www.neurotechnology.com/verifinger.html

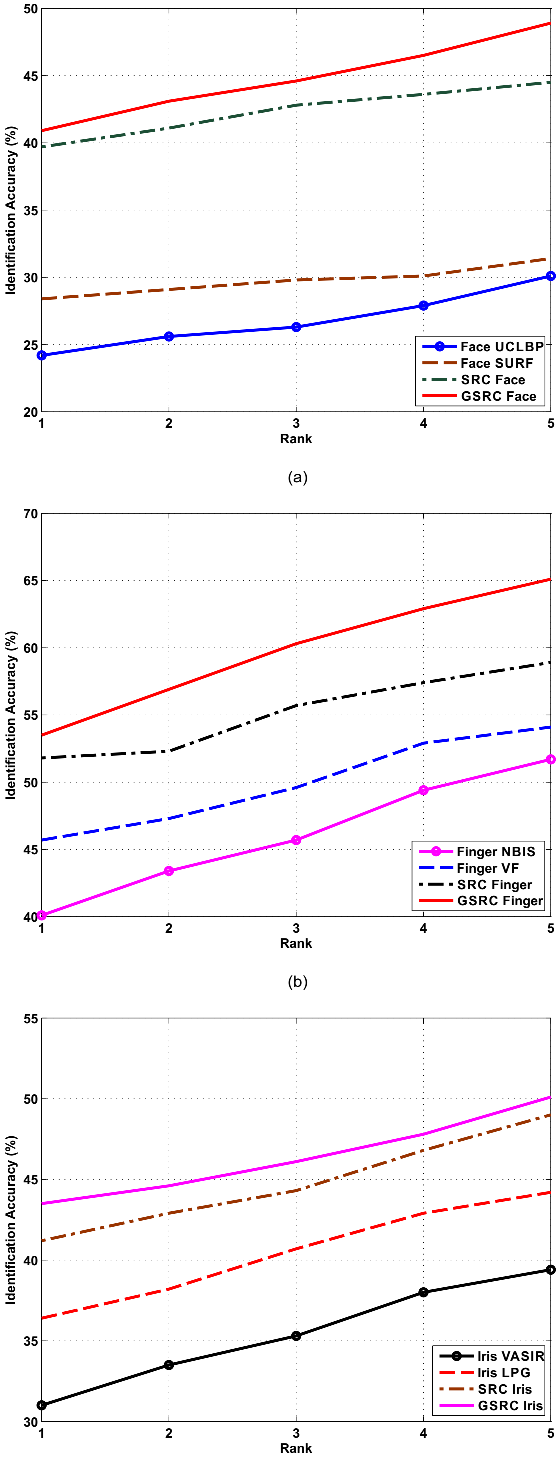


(a)

(b)

(c)

Figure 3-4: CMC curves on the WVU multimodal database: individual features, SRC and GSRC on (a) face, (b) fingerprint and (c) iris.



(a)

(b)

(c)

Figure 3-5: CMC curves on the LEA multimodal database: individual features, SRC and GSRC on (a) face, (b) fingerprint,and (c) iris.

Table 3.3: Rank-1 identification accuracy (%) with fusion of multiple modalities and multiple features (SRC and GSRC) on the WVU and LEA databases.

Modality	**Fusion Algorithm**	**WVU**	**LEA**
Face and Iris	SRC	93.9	45.3
	GSRC	94.6	47.4
Face and Finger	SRC	94.4	52.1
	GSRC	95.3	55.8
Iris and Finger	SRC	95.6	52.5
	GSRC	95.9	55.1
Face, Finger and Iris	Sum Rule (score level)	95.0	52.6
	Context Switching [15]	95.8	55.8
	SRC	95.1	54.6
	GSRC	99.1	62.3

for each modality. UCLBP [198] and SURF [9] features are considered for face, VASIR and LPG are considered for iris, and NBIS and VF are used for fingerprint. It is observed that the fingerprint and iris modalities perform better than face on both databases; however, fingerprint features outperform the iris modality on the LEA database, possibly denoting the higher reliability of fingerprint modality when data is captured in unconstrained, real-world conditions. It is also observed that no single modality or feature offers the best performance, particularly on the LEA database, clearly motivating the requirement for a multimodal multi-feature recognition algorithm.

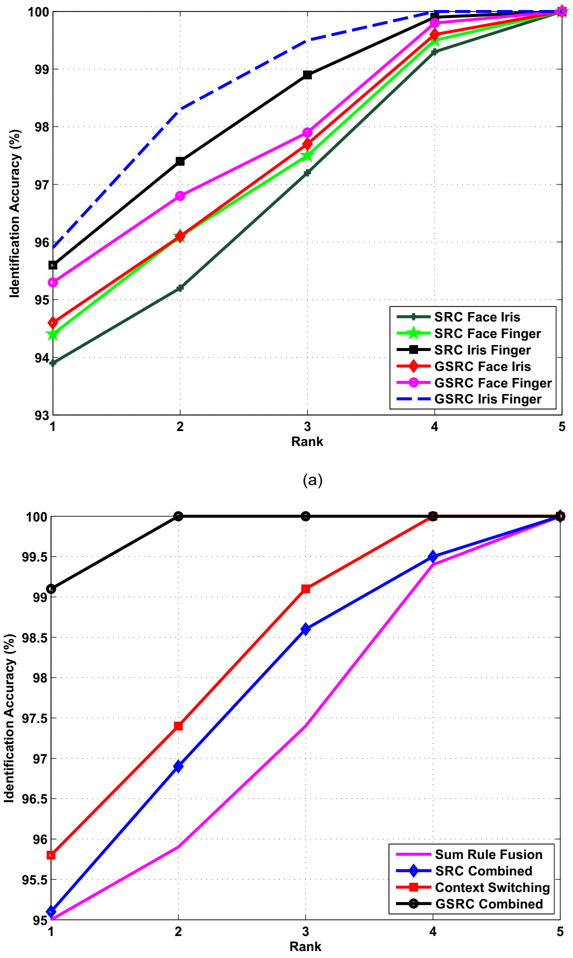
- **Single-modality multi-feature**: These experiments are performed to evaluate the difference in performance of the traditional SRC algorithm with the proposed GSRC algorithm when multiple features are considered for each modality. In accordance with existing literature, decisions from SRC classifiers operating on individual features from the same modality are combined using sum rule fusion at the match score level [94]. From Figures 3-4 and 3-5, it is evident that the GSRC algorithm performs better than the SRC based algorithm on both the databases. However, both of them improve the performance considerably compared to a single feature. Further, we perform McNemar test and at 95% confidence, using the rank-1 results on LEA database (in Table 2), it is observed that SRC and GSRC are statistically different.

- **Two-modality multi-feature**: These experiments are performed to evaluate the performance of the traditional SRC algorithm and the proposed GSRC algorithm when both features for two modalities are combined at a time. Similar to the above, sum rule fusion is performed at match score level to obtain SRC decisions. As shown in Table 3 and Figures 3-6(a) and 3-7(a), the proposed GSRC algorithm outperforms SRC on both the databases. The difference in performance of SRC and GSRC is higher, especially at lower ranks. We have also observed that if the data is of higher quality then GSRC can offer better performance improvement.

- **Multimodality multi-feature**: These experiments are performed to evaluate the performance of the proposed GSRC when all the modalities and features are utilized, i.e, three modalities and two features each. The results are presented in Figure 3-6(b) and Figure 3-7(b). On the WVU database, traditional SRC achieves 95.1% rank-1 accuracy whereas the GSRC algorithm achieves 99.1% rank-1 performance. On the LEA database, traditional SRC achieves 54.6% rank-1 accuracy whereas the GSRC algorithm achieves 62.3% performance. This demonstrates that the GSRC algorithm is more robust since the LEA database consists of images with lower quality and unconstrained pose, illumination, and expression.

- **Comparison with existing algorithms**: For the WVU database, at rank 1, the proposed algorithm achieves an identification accuracy of 99.1%, whereas the second best performance of 95.8% is obtained using the context switching algorithm [15]. GSRC also achieves 100% accuracy at rank 2 for the WVU database, whereas the existing state-of-the-art achieves it at rank 4. On the LEA database, the context switching algorithm outperforms both sum rule fusion and traditional SRC with an identification performance of 55.8% whereas the GSRC algorithm performs 6.5% better and achieves 62.3% rank-1 accuracy. It is to be noted that an important characteristic of the context switching algorithm is that it is able to handle missing data. In case of the LEA database, there are over 75% cases which have at least one missing modality. In our experiments, we observe that the GSRC algorithm can also handle the missing data cases and perform better recognition.

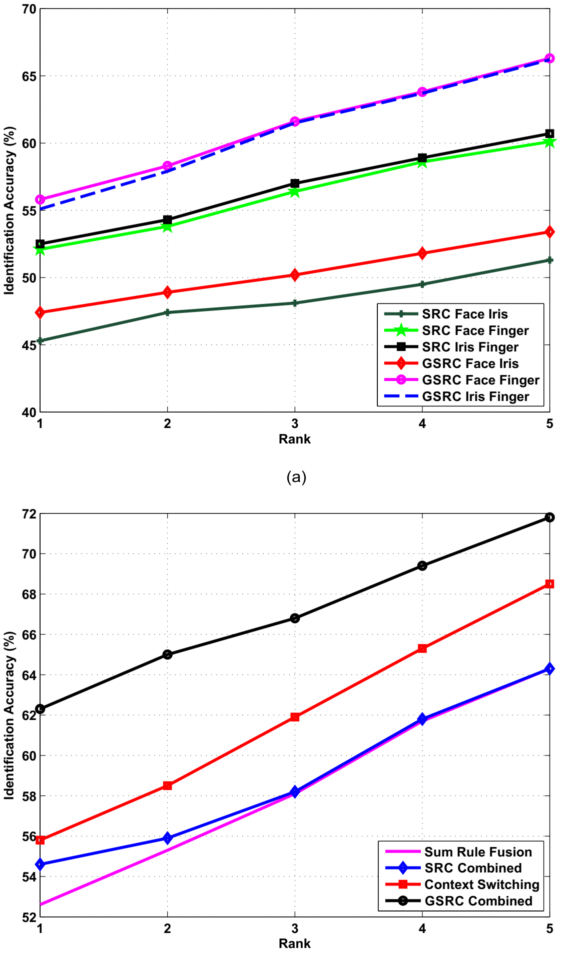
- **Performance with missing data**: To further accentuate the effect of partial data (missing



(a)

(b)

Figure 3-6: CMC curves for identification on the WVU multimodal database. Comparing the performance of (a) traditional SRC with the proposed GSRC when two modalities are considered at a time and (b) GSRC, SRC, sum rule, and context switching algorithms when all three modalities are considered at a time.



(a)

(b)

Figure 3-7: CMC curves for identification on the LEA multimodal database. Comparing the performance of (a) traditional SRC with the proposed GSRC when two modalities are considered at a time and (b) GSRC, SRC, sum rule, and context switching algorithms when all three modalities are considered at a time.

Table 3.4: The impact of feature characteristics on the performance of GSRC. ^{1} = original, ^{2} = normalized.

Classifier	LEA		WVU	
	UCLBP + SURF ^{1}	UCLBP + SURF ^{2}	UCLBP + SURF ^{1}	UCLBP + SURF ^{2}
SRC	29.3	39.7	80.1	82.3
GSRC	30.2	40.9	80.8	83.7

Table 3.5: The impact of gallery size on the performance of GSRC.

Classifier	WVU		LEA	
	Gallery = 2	Gallery = 3	Gallery = 2	Gallery = 3
SRC	45.2	54.6	88.4	95.1
GSRC	53.9	62.3	94.7	99.1

modality scenario), an experiment on LEA database is conducted in which for each multimodal probe, sample from one modality is either randomly removed or not available in the database (75% of the cases), i.e. this experiment mimics the scenario when all probe samples have at least one modality missing. We have observed that in such scenario, GSRC yields an accuracy of 61.8% whereas the next best performance of 55.4% is achieved with the existing context switching algorithm and SRC and Sum Rule yield an accuracy of 52.9% and 50.2%. This showcases that the proposed GSRC algorithm is able to handle missing features better compared to other algorithms.

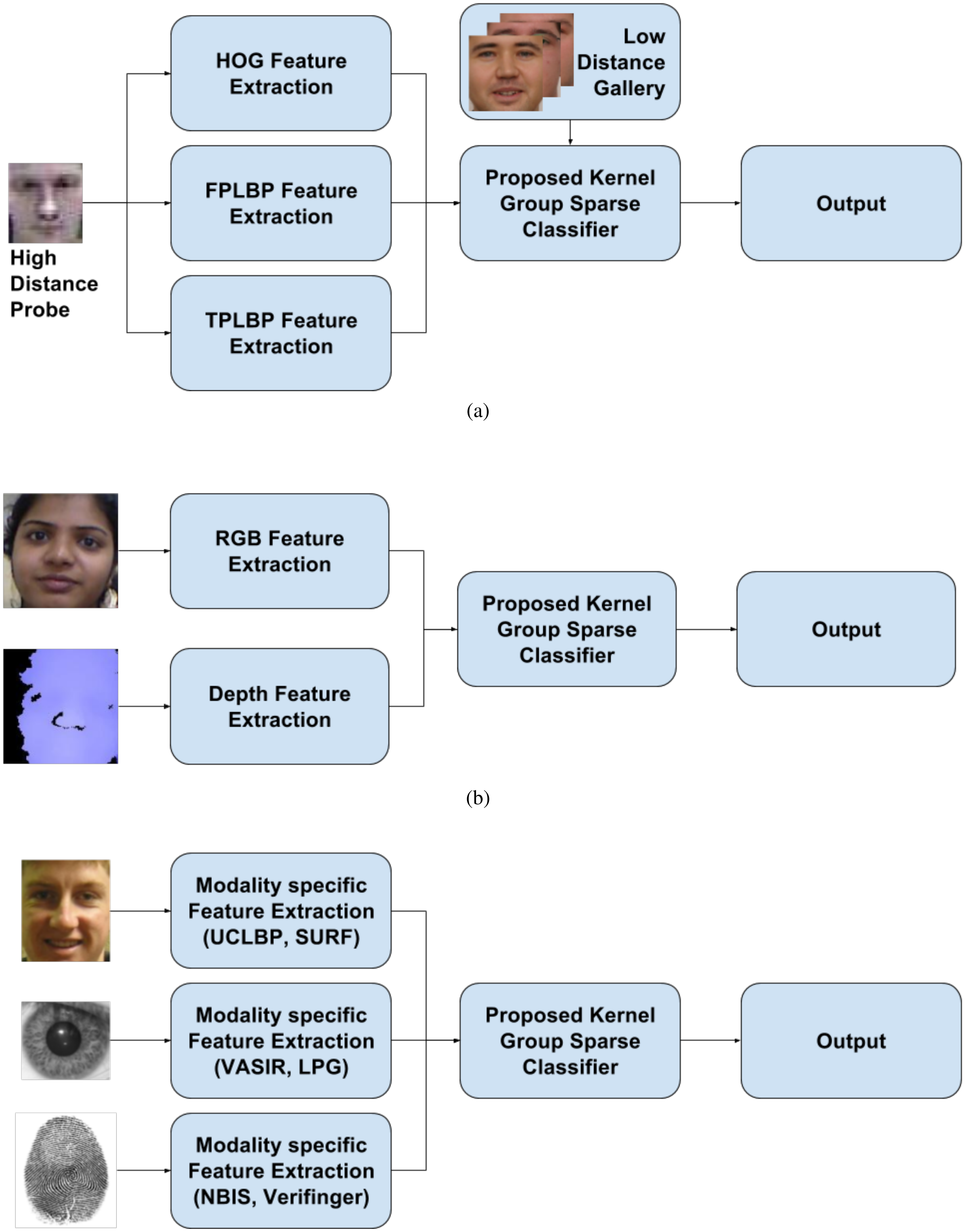
- **Impact of feature characteristics**: We observe that when a weak feature of higher feature length is combined with a strong feature of significantly lower feature length, the performance of GSRC suffers. Revelant results are included in Table 3.4. Therefore, it is important to exercise caution with utilizing features with variable lengths without normalization. However, when feature lengths are normalized and two weak features are combined, GSRC performs much better compared to the SRC algorithm and boosts the combined performance by a large margin, e.g., in the case of the face modality. It is also observed that when the features are similar in nature (Gabor features in case of iris modality, minutiae features in case of fingerprint modality), GSRC improves classification performance by a relatively smaller margin.

- **Dependency on gallery/training data**: Any sparse representation based scheme requires labeled samples in order to compute efficient sparse representations and retain discriminative capabilities. However, we observe that when the data points available per class is reduced, the performance of the SRC algorithm suffers much more than the GSRC algorithm. Revelant results are included in Table 3.5.We also observe that if the feature length is increased, GSRC can readily utilize the additional information to improve performance higher than the SRC algorithm.

- **Computational time**: The computational complexity of the proposed algorithm is primarily dependent on the the $l_{2,1}$-minimization and has the overall time complexity of $O\left(NTlog(NT)\right)$. The running time of both traditional SRC and the proposed GSRC algorithm is compared on an Intel machine with Core i7 3.40 GHz quad core processor, 16 GB of RAM, and MATLAB programming environment. SRC requires 0.11 seconds to process one probe image and the GSRC algorithm requires 0.51 seconds to process one multi-modal probe with all three modalities present. Sum rule fusion at the score level requires 0.01 second and context switching algorithm requires 0.27 seconds. Note that this time does not include time spent in acquisition, preprocessing or feature extraction which are common to both approaches and contribute equally to the computational time of both classifiers.

3.5 KGSRC: Case Studies

In this research, we apply the proposed KGSRC algorithm to three challenging biometrics problems, (1) cross-distance face recognition, (2) RGB-D face recognition, and (3) multi-modal biometric recognition, to assess its performance. The flow diagrams for each of these problems is presented in Figure 3-8. For a given probe sample, individual representations are extracted using different feature extractors, modalities, or spectrums as applicable. A sparse representation for the probe image is computed over the features extracted from the gallery images using the proposed KGSRC algorithm. The residual error for each subject is compared and the probe is assigned to the subject/class with the minimum residual error. In each case, face detection is performed using the Viola Jones detector [207] before processing an image during either training or testing and only the cropped face image is used. To populate our kernel list we have used the linear, spline, wave, sigmoid, radial basis function, polynomial, multiquadratic, rational quadratic, and laplacian ker-


HOG Feature Extraction
FPLBP Feature Extraction
TPLBP Feature Extraction
High Distance Probe
Low Distance Gallery
Proposed Kernel Group Sparse Classifier
Output
RGB Feature Extraction
Depth Feature Extraction
Proposed Kernel Group Sparse Classifier
Output
Modality specific Feature Extraction (UCLBP, SURF)
Modality specific Feature Extraction (VASIR, LPG)
Modality specific Feature Extraction (NBIS, Verifinger)
Proposed Kernel Group Sparse Classifier
Output

(a)

(b)

(c)

Figure 3-8: Flow diagrams for each of the three biometric problems that we use to evaluate the proposed KGSRC algorithm. (a) Cross-distance face recognition (b) RGB-D face recognition (c) Multi-modal biometric recognition

nel classes along with correspondingly appropriate parameter ranges. For the purpose of feature extraction, we explore three existing feature descriptors: TPLBP [215], FPLBP [215], and HOG [35]. These descriptors are selected since LBP and HOG based descriptors have been successfully applied for RGB-D face recognition in existing literature and have been demonstrated to achieve high performance.

3.5.1 Cross-distance Face Recognition

Cross-distance face recognition is a challenging problem which occurs frequently when surveillance footage needs to be used for face recognition. In such a case, the face image is captured at a high distance from a fixed camera position that inherently imparts the problems of pose and illumination variation along with low-resolution. These images are then usually matched with mugshot images which are of high quality. To use the KGSRC algorithm in this scenario, we first perform face detection and crop the face image only from the background. We then utilize three different feature extractors and perform feature level fusion using the KGSRC algorithm. The training data is used as gallery for each subject, to learn the ideal kernel parameters, and to learn the sparse representation coefficients. The same process of extracting multiple features is employed for a given probe image at runtime and the KGSRC algorithm computes a ranked list of subjects ordered by increasing residual error.

3.5.2 RGB-D Face Recognition

RGB-D face recognition focuses on recognizing faces when they are captured in the form of a color image (RGB) along with a corresponding depth map (D). The color image and depth map pair is called as one RGB-D image. The methods of acquisition of the depth map varies from device to device but the Microsoft Kinect device uses an infrared laser projector combined with a monochrome CMOS sensor. Two public databases [66, 166] containing RGB-D face images captured using the Kinect sesnor are used in this research. Given a RGB-D face image, registration of the faces is performed to increase the correspondence between the pixels of the depth map and the color image. Then, face detection is performed using the color face image and the same bounding box is used for the depth map as well. The color face image and the depth map act as independent feature sources and features can be extracted from both. Each subject's gallery images are used to train the proposed KGSRC algorithm and learn the corresponding sparse representation. Given a probe image, features are extracted from both the color image and the depth map separately and matched with the gallery using the KGSRC algorithm and a ranked list of subjects is obtained by sorting them by ascending order of the residual error.

3.5.3 Multi-modal Biometric Recognition

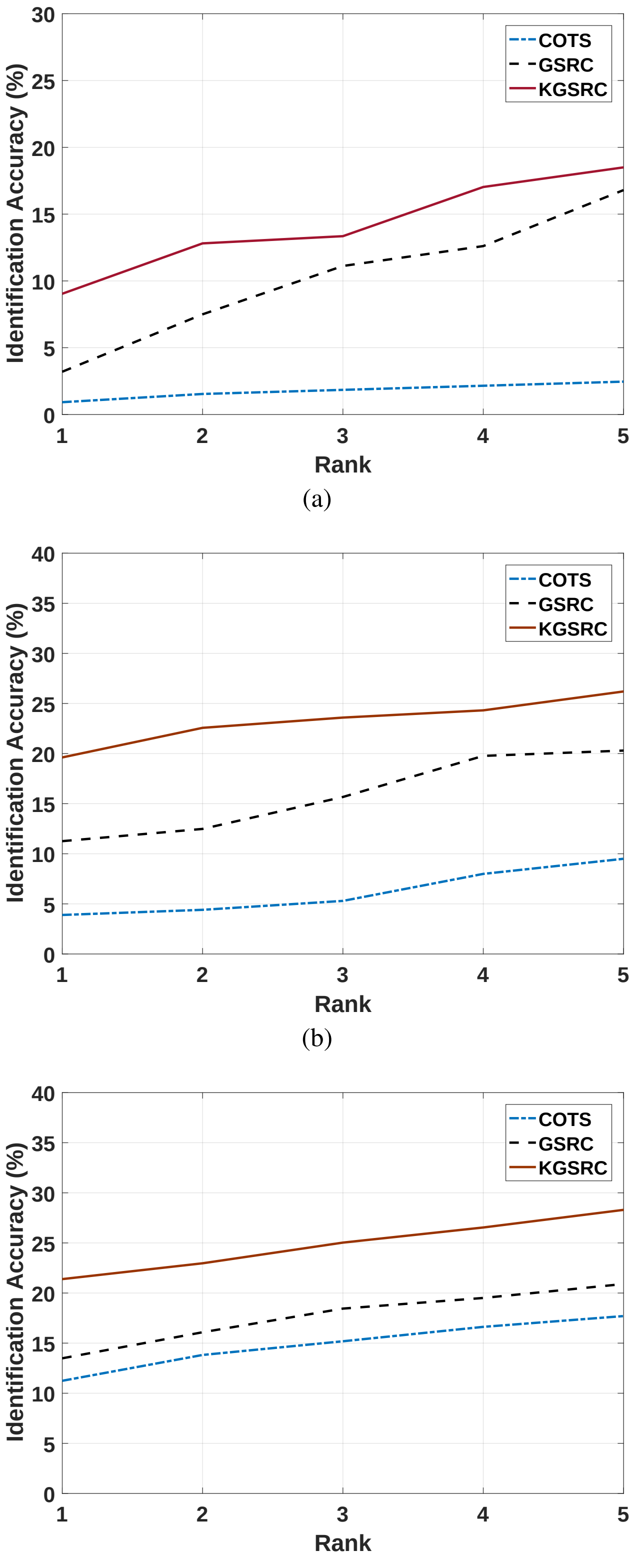
In multi-modal biometric recognition, multiple biometric modalities are used to determine the identity of a given individual. The most popular biometric modalities are face, fingerprint, and iris. In this scenario, each modality acts as an independent feature source. Appropriate pre-processing is applied to the image from each modality and then as shown in Figure 3-8c features are extracted using a feature extractor relevant to the modality. Then, the corresponding features are used for training the KGSRC algorithm. A given probe image may be identified on the basis of any combination of modalities (e.g. only face, only iris, or face and iris or all the modalities) since the algorithm can inherently handle the missing data problem. The residual error is computed for each subject and a ranked list of candidate matches is produced based on assigning the highest rank to the subject with the least residual error.

3.6 KGSRC: Experimental Results and Analysis

The following subsections present experimental results pertaining to each of the case studies including database and protocol details for each of the problems pertaining to each of the problems:

3.6.1 Cross-distance Face Recognition in Surveillance

The performance of the proposed algorithm is evaluated on the publicly available SCFace database [68]. It contains 4,160 human face images for 130 subjects in both visible and near-infrared spectrums. The images are captured in uncontrolled indoor environments using five video surveillance cameras of varying quality and at variable distances. The database also provides five good quality close-distance face images per subject for the purpose of enrollment/gallery data. These mugshots

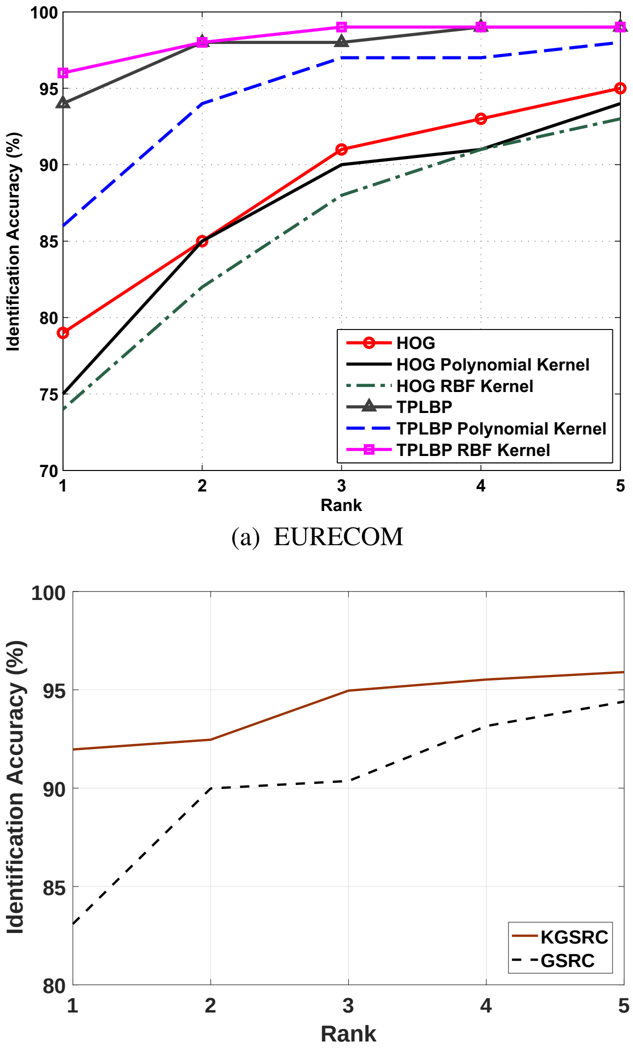


(a)

(b)

(c)

Figure 3-9: Identification results on the SCFace database at three different distances of the probe image: (a) Distance 1 = 4.2 metres, (b) Distance 2 = 2.6 metres, and (c) Distance 3 = 1.0 metres. The matching is performed between a high distance probe and a low distance (high resolution) gallery. For both the KGSRC and GSRC algorithms, the best feature combination is showcased in the CMC curve.



(a) EURECOM

(b) IIIT-D

Figure 3-10: Identification results on the two RGB-D databases when compared to GSRC. For both the KGSRC and GSRC algorithms, the best feature combination is showcased in the CMC. We see that kernelization using the proposed algorithm improves the performance.

Table 3.6: Identification results at rank 5 on the SCFace database. Distance 1 is the furthest and distance 3 is the closest from the camera.

Algorithm	Features	Rank 5 Identification Accuracy (%)		
		Distance 1	Distance 2	Distance 3
GSRC	HOG + FPLBP	16.8	17.4	12.5
	HOG + TPLBP	15.7	19.7	20.9
	TPLBP + FPLBP	13.7	13.8	15.1
	HOG + FPLBP + TPLBP	16.7	20.3	17.2
COTS		2.5	9.5	17.7
Proposed KGSRC	HOG + FPLBP	18.3	**26.2**	27.5
	HOG + TPLBP	12.9	23.8	**28.3**
	TPLBP + FPLBP	11.1	10.0	15.1
	HOG + FPLBP + TPLBP	**18.5**	25.2	27.5

are captured using a high-quality photo camera. In order to conduct our experiments, we utilize the most frontal three good quality faces per subject as gallery. Then, we perform three experiments for each of the three different distances in the database: distance 1 is furthest from the sensor, whereas distance 3 is the closest to the sensor. All the images available for a subject for each distance are used as probe. Therefore, there are a total of 390 gallery and 650 probe images per distance. Distances 1, 2, and 3, denote distances of 4.20, 2.60, and 1.00 metres from the camera, respectively. Other capture conditions in the database are also reflective of the real-world scenario, i.e., the camera is placed slightly above the subject's head. The database encompasses a highly challenging recognition scenario. We compare the performance of the proposed algorithm to a state-of-the-art COTS face recognition system, as well as with traditional GSRC approach. The results, in the form of rank 5 identification accuracy and CMC curves, are presented in Table 3.6 and Figure 3-9, respectively.

We observe that the proposed KGSRC algorithm offers substantial improvements over the COTS algorithm, achieving a best case performance of 18.5%, 26.2%, and 28.3% rank 5 identification accuracy at distances 1, 2, and 3, respectively. In comparison, the COTS approach is only able to achieve 1.5%, 9.5%, and 17.7% accuracy at distances 1, 2, and 3 respectively using the same protocol. As opposed to the non-kernelized version (GSRC), the proposed KGSRC algorithm offers an improvement of 1.7%, 5.9%, and 7.38% at distances 1, 2, and 3, respectively. This implies that a total of 11, 38, and 48 probe images that are misclassified by the GSRC are identified correctly by the proposed algorithm at distances 1, 2, and 3, respectively.

Table 3.7: Identification results at rank 5 on both Eurecom and IIIT-D RGB-D databases. We use the existing protocol and partitions [66] to enable direct comparison. $*$ Results have been taken from [66].

	Algorithm	Rank-5 Accuracy (%)	
		Eurecom	IIIT-D
Existing	Goswami *et al.* $*$[66]	98.5 ± 1.6	95.3 ± 1.7
	3D-PCA $*$	94.1 ± 2.7	83.4 ± 2.1
	FPLBP $*$	94.3 ± 1.4	85.0 ± 0.7
	HOG $*$	89.5 ± 0.8	75.1 ± 0.7
	SIFT $*$	83.8 ± 2.1	50.1 ± 1.4
	Sparse $*$	84.8 ± 1.7	87.2 ± 1.9
	HOG + GSRC	95.4 ± 1.7	89.9 ± 2.8
	TPLBP + GSRC	97.2 ± 1.2	94.4 ± 1.5
Proposed	HOG + KGSRC	96.7 ± 1.5	91.8 ± 2.8
	TPLBP + KGSRC	**98.8** ± **1.2**	**95.9** ± **1.6**

3.6.2 RGB-D Face Recognition

The performance of the proposed algorithm is evaluated on two publicly available RGB-D databases: the EURECOM database [166] and the IIIT-D RGB-D database [66]. First, we present a brief overview of each of these databases and explain the experimental protocol. Next, we present identification results on these databases using the proposed algorithm and also present a comparison with existing state-of-the-art algorithms on these databases.

The EURECOM database [166] consists of RGB-D images of 52 individuals captured in two different sessions. The database consists of variations in illumination and expression, and consists of a total of 936 images. In order to compare results with the state-of-the-art algorithms, we follow the existing protocol [66]. The entirety of the data is partitioned such that data for 31 subjects is in the testing partition and the remaining 21 subjects are part of the training partition. This partitioning is performed five times for cross validation. We report the mean accuracies at each rank in the form of a CMC curve in Figure 3-10a. Mean accuracy and standard deviation at rank 5 are presented in Table 3.7.

The IIIT-D Kinect RGB-D face database [66] consists of RGB-D images of 106 individuals (4605 images), captured in two different sessions. Similar to the case of the EURECOM database, existing protocol [66] is used to facilitate comparative results. Each partition contains training data pertaining to 42 subjects and testing data pertaining to 64 subjects. We report the mean accuracies at each rank in the form of a CMC curve in Figure 3-10b. Mean accuracy and standard deviation at rank 5 are presented in Table 3.7. Along with demonstrating the performance of the proposed algorithm with different feature extractors on the two databases, we have compared the performance with existing state-of-the-art algorithms on the two databases. These algorithms are Goswami et al. [66], 3D-PCA, FPLBP, HOG, SIFT, and sparse classification.

On both the databases, the proposed KGSRC algorithm performs the best, achieving 98.8% and 95.9% rank 5 identification accuracies on the Eurecom and IIIT-D RGB-D databases, respectively. It outperforms its non-kernel counterpart by 1.6% on the IIIT-D database and 1.5% on the EURECOM database. It is also able to outperform the previous best performing algorithm by 0.3% and 0.6%, achieving results that are comparable to the state-of-the-art on the two databases. We also observe that the KGSRC algorithm performs consistently better with TPLBP than with HOG. In the current implementation, HOG is a local feature descriptor of a smaller size (256) whereas the TPLBP descriptor provides a larger feature representation of size 3,072. This can be attributed to the nature of sparse classification algorithms that perform better if over-complete representations are provided as input.

3.6.3 Multi-modal Biometric Recognition

The WVU multi-modal database [34] consists of data pertaining to multiple biometric modalities for 270 individuals. In this research, we focus on three popular biometric modalities among these: iris, face, and fingerprint. Since not every modality is available for all individuals, the database incorporates the missing data problem. 231, 270, and 316 subjects have data available for the iris, fingerprint, and face modalities, respectively. Data pertaining to 108 subjects (40%) is utilized for training and data for the remaining 162 subjects (60%) is utilized for testing. Three images are used as gallery and training and the remaining images are probes.

The results of the evaluation are presented in Table 3.8. We observe that the proposed KGSRC algorithm achieves the best performance among all the evaluated algorithms. We notice an improvement of 0.6% on the final fusion of all three modalities and is able to further improve upon the already near perfect accuracy achieved by the group sparse classifier. Among two modality fusion experiments, we see consistent improvement of about 2-3% when the group sparse clas-

Table 3.8: Rank-1 identification accuracy (%) with fusion of multiple modalities and multiple features (SRC, GSRC, and KGSRC) on the WVU database.

Modality	**Features**		**WVU**
Face	Individual	UCLBP	75.4
		SURF	79.1
	Fusion	SRC	82.3
		GSRC	83.7
		KGSRC	**85.2**
Iris	Individual	Vasir	85.0
		LPG	90.5
	Fusion	SRC	92.9
		GSRC	93.5
		KGSRC	**96.1**
Finger	Individual	NBIS	85.9
		VeriFinger	90.7
	Fusion	SRC	92.6
		GSRC	93.1
		KGSRC	**95.8**
Face and Iris	Fusion	SRC	93.9
		GSRC	94.6
		KGSRC	**97.4**
Face and Finger	Fusion	SRC	94.4
		GSRC	95.3
		KGSRC	**97.1**
Iris and Finger	Fusion	SRC	95.6
		GSRC	95.9
		KGSRC	**98.2**
Face, Finger and Iris	Fusion	Sum Rule (score level)	95.0
		Context Switching [15]	95.8
		SRC	95.1
		GSRC	99.1
		KGSRC	**99.7**

sifier already achieves over 95% accuracy on the database, leaving little room for improvement. Therefore, we can assess that the kernelization of the features before matching helps in improving the discriminative capability of the classifier and results in a more robust feature fusion methodology. The algorithm is able to successfully classify samples that are not separable in the relatively lower dimensional space of the non-kernel group sparse classifier by projecting them to a higher dimensional space using the automatically determined kernel function. We also see experimental support for the efficacy of the proposed automatic kernel and parameter selection metric since the choice of kernel and parameter can greatly influence the success of any kernel based approach.

3.7 Summary

Biometrics is a challenging problem due to various covariates such as pose, illumination, and expression, which can adversely influence recognition performance. However, utilizing multiple features to represent each sample can provide robustness and enhance the accuracy of recognition algorithms. In this research, we present the group sparse representation based classifier for multimodal multi-feature biometric recognition. The proposed algorithm operates on the feature vectors obtained from different modalities/descriptors and perform recognition via feature level fusion and classification. Results on two multimodal databases showcase the efficiency of the proposed algorithm in comparison to existing state-of-the-art algorithms. The GSRC algorithm is able to encode the complementary information obtained using different modalities and features to perform accurate identification in unconstrained scenarios. We also present the KGSRC which is an extension of the GSRC framework with kernelization. Using the training data, we not only learn the coefficients for sparse coding to generate the most discriminative representation of data, but also learn the appropriate kernel function and associated parameter values automatically. Using these case studies on cross-distance face recognition, RGB-D face recognition, and multimodal biometric fusion, we observe that the KGSRC algorithm with the kernel selection methodology, obtains comparable results to the existing state-of-the-art.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Face Verification via Learned Representation on Feature-Rich Video Frames

Video face recognition has become highly significant in surveillance scenarios. For example, more than 80,000 people were identified and verified during the 2008 Beijing Olympics with the help of face recognition in videos [1]. With advancements in technology, video capturing devices are accessible to a large number of people in the form of portable electronic devices such as phones and tablets. In unconstrained scenarios, videos captured by such devices may also be used by law enforcement agencies. Therefore, there is a high motivation to utilize video data to perform accurate face recognition. Figure 4-1 shows frames from four video clips in which the face regions have been detected and cropped. While a single frame from a video can only capture limited information, multiple frames capture a lot of information about the face pertaining to its appearance under the effect of common covariates such as pose, illumination, and expression. By utilizing the large variety of information present in a video, a robust and comprehensive representation of a face can be extracted and accuracy can be improved.

Video face recognition has been extensively studied and several algorithms have been proposed. Table 4.1 provides a review of some of the algorithms along with the summary of results reported on popular video face recognition databases. Video face recognition algorithms can broadly be classified into two types: (a) set-based and (b) sequence-based [92]. The set-based

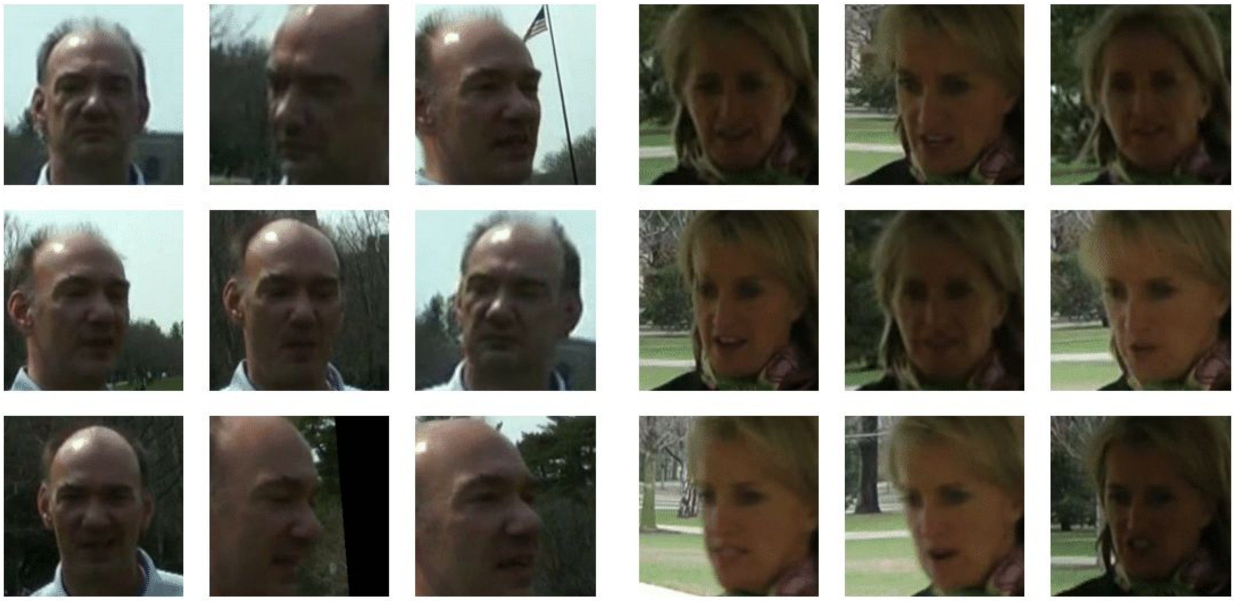


Figure 4-1: A subset of frames illustrating the amount of information present in a video. A single video can capture a subject's face under different pose, expression, and illumination variations. While some frames can be highly useful for face recognition, others can be detrimental to performance. Images are frames from the PaSC database [12].

approaches consider a video as a set of images (frames) which are then modeled and matched using a variety of methodologies. These approaches may not utilize the temporal information contained in the video, i.e. the order of frames in the original video may not matter. On the other hand, sequence-based approaches are specifically designed to utilize temporal information of the video. These approaches model the video as a sequence of images and apply sequence classification techniques for recognition. Some of the recent techniques utilize large image dictionaries to characterize videos [29, 74], while some others have focused on metric learning based approaches [91] or deep learning based approaches [26, 222]. For comparison, the results are generally reported on benchmark databases such as the Honda UCSD database [98], YTF [114], and recently developed PaSC database [12].

Authors	Algorithm	Database	Verification Accuracy
Wolf *et al.*, 2011 [114][1]	Matched background similarity L2 mean with LBP	YTF [114]	76.4%
Wolf and Levy, 2013 [216][1]	SVM-Minus similarity score with background similarity		78.9%
Li *et al.*, 2013 [72][1]	Probabilistic elastic matching		79.1%

Cui *et al.*, 2013 [234][2]	Spatial-temporal face region descriptor + Pairwise-constrained multiple metric learning		79.5%
Mendez-Vazquez *et al.*, 2013 [73][2]	Volume structured ordinal features		79.7%
Bhatt *et al.*, 2014 [74][1]	Clustering based re-ranking and fusion		80.7%
Hu *et al.*, 2014 [82][1]	Large margin multi-metric learning for face and kinship verification in the wild		81.3%
Hu *et al.*, 2014 [91][1]	Discriminative deep metric learning		82.3%
Taigman *et al.*, 2014 [222][1]	Nine-layer deep network		91.4 % (unrestricted)
Wang *et al.*, 2015 [211][1]	Discriminant Analysis on Riemannian manifold of Gaussian distributions		73.01 AUC
Khan *et al.*, 2015 [99][1]	Adaptive Sparse Dictionary		82.9%
Li *et al.*, 2015 [118][1]	Eigen-PEP for video face recognition		84.8%
Li *et al.*, 2015 [116][1]	Hierarchical-PEP for video face recognition		87.0%
Sun *et al.*, 2015 [193][1]	Semi-supervised convolutional neural network		93.2% (unrestricted)
Schroff *et al.*, 2015 [183][1]	Unified embedding learned using deep CNN		95.1% (unrestricted)
Parkhi *et al.*, 2015 [160][1]	Eleven-layer deep convolutional neural network with triplet loss based face embedding		97.3% (unrestricted)
Ding and Tao, 2016 [42][1]	Ensemble of Deep Convolutional Neural Networks		94.96% (unrestricted)
Yang *et al.*, 2016 [227][1]	GoogLeNet [196] features with aggregation		95.5% (unrestricted)
Tran *et al.*, 2016 [201][1]	3D Morphable Face Models regressed using a CNN		88.8% (unrestricted)
Ranjan *et al.*, 2018 [168]	Crystal loss and quality pooling		96.08% (unrestricted)

Beveridge *et al.*, 2013 [12][1]	Local region principal component analysis		8% (handheld) 10% (control)
Wang *et al.*, 2015 [211][1]	Discriminant Analysis on Riemannian manifold		18.3% (handheld) 18.7% (control)
Li *et al.*, 2015 [116][1]	Hierarchical-PEP for video face recognition		30.7%

PaSC [12]

Huang *et al.*, 2015 [88][1]	Projection Metric Learning on Grassmann Manifold		43.9% (handheld) 43.6% (control)
Huang *et al.*, 2015 [11][1]	Hybrid Euclidean-and-Riemannian Metric Learning		59% (handheld) 58% (control)
Ding and Tao, 2016 [42][1]	Ensemble of Deep Convolutional Neural Networks		95.9% (handheld-unrestricted) 96.2% (control-unrestricted)
Ding and Tao *et al.*, 2017 [43][1]	Pose-invariance with homography-based normalization		60.4% (handheld) 69.1% (control)
Rao *et al.*, 2017 [170][1]	Attention-aware Deep Reinforcement Learning		93.8% (handheld) 95.7% (control)
Wang *et al.*, 2017 [212][1]	Discriminative Covariance Oriented Representation Learning		55.7% (handheld) 56.4% (control)
Ding and Tao, 2018 [44][1]	Trunk-branch ensemble convolutional neural networks + batch normalization		96.1% (handheld-unrestricted) 97.8% (control-unrestricted)
Goswami et al. [62][1]	Memorability based frame selection and deep learning	PaSC	89% (handheld), 94% (controlled)
		YTF	88.6% (unrestricted)
Proposed[1]	Feature-richness based frame selection and deep learning (joint learning in autoencoder with sparse and low rank DBM)	YTF [114]	93.4%, 95.4% (unrestricted)
		PaSC [12]	93.1% (handheld), 97.2% (handheld-unrestricted), 95.9% (control), 98.1% (control-unrestricted)

Table 4.1: Review of selected papers on video face recognition that have shown results on the YTF and PaSC benchmark face video databases. Results marked unrestricted denote algorithms that have used external training data during training. The algorithms follow the standard experimental protocol during testing for both databases to facilitate comparison. [1]*denotes set based algorithms*, [2]*denotes sequence based algorithms*.

As shown in Table 4.1, existing algorithms have attained high performance on YTF [114]. However, the protocol of this databases generally require reporting the results at EER [175]. From implementation perspective, the algorithms are required to minimize False Accept Rate (FAR) or False Reject Rate (FRR). However, lower EER does not necessarily mean low FAR or FRR. Figure 4-2 illustrates the performance of some of the existing algorithms on the YTF [114]. It

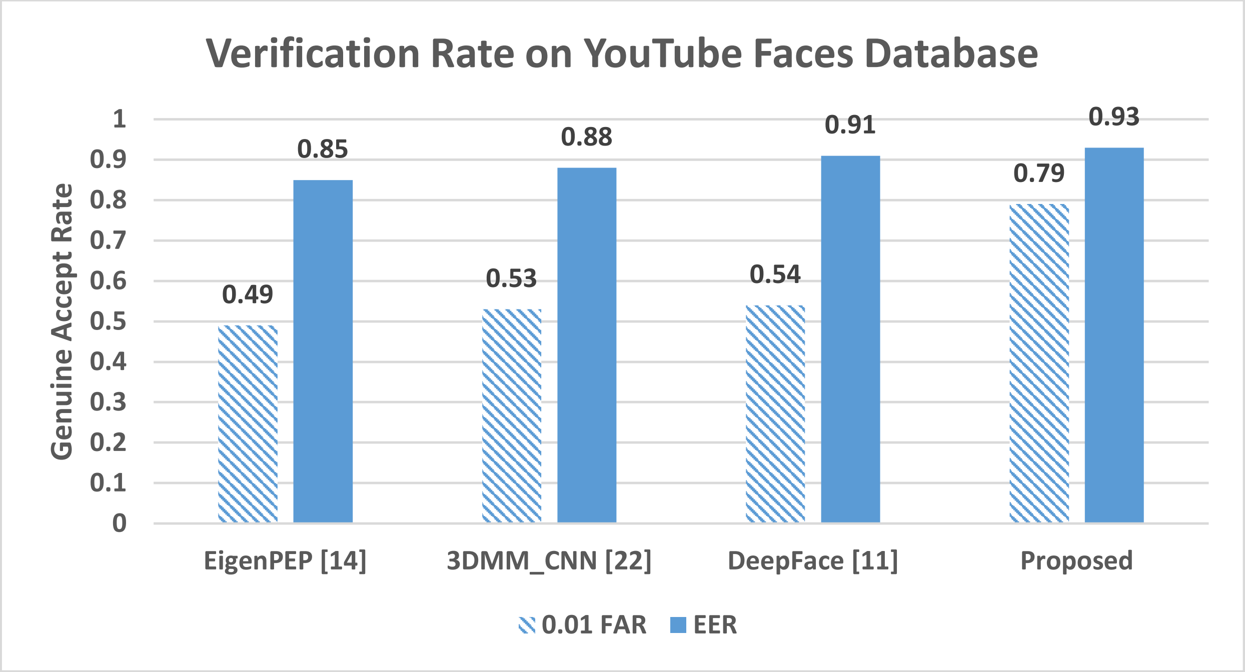


Figure 4-2: Summarizing the performance of some of the best performing face verification algorithms on the YTF [114]. It is evident that there is a huge gap in the performance at low false accept rates as compared to performance at EER. We showcase that the proposed algorithm performs well even at a low false accept rate.

is observed that these algorithms attain very high accuracies at equal error rate, however, their performance at lower false accept rates is significantly lower. For example, DeepFace [222] yields over 91% verification accuracy at EER but only 54.1% at 1% FAR. For many security related applications, such as video surveillance, it is desirable to achieve high verification performance while minimizing the false accept rates. Therefore, it is our assertion that there is a significant scope of improvement in the performance of video face recognition and additional research is required, especially focusing at lower false accept rates.

In general, video face verification involves matching using all the frames present in two videos. However, not all frames are equally informative and some frames might suffer from low image quality or extreme variations due to pose, expression, and illumination. Due to the presence of these covariates of face recognition, some frames may affect the inter-class and intra-class variations. In other words, it is highly probable that features extracted from such a frame might lead to incorrect results. Therefore, it is important to select and utilize the high information content in a video carefully and efficiently which makes video data more challenging as well as rewarding

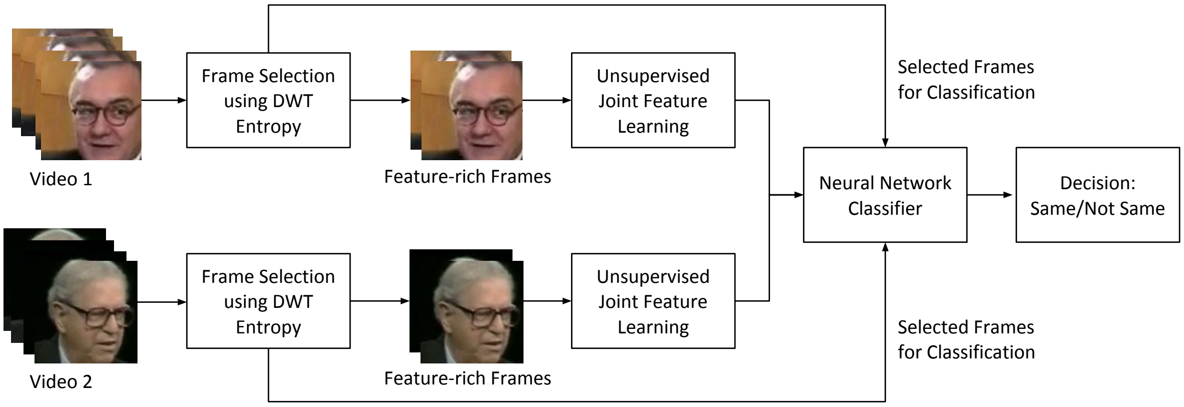

Video 1
Frame Selection using DWT Entropy
Feature-rich Frames
Unsupervised Joint Feature Learning
Selected Frames for Classification
Neural Network Classifier
Decision: Same/Not Same
Video 2
Frame Selection using DWT Entropy
Feature-rich Frames
Unsupervised Joint Feature Learning
Selected Frames for Classification

Figure 4-3: Illustrating the steps involved in the proposed face recognition algorithm.

for face recognition. To address some of these limitations and to improve overall performance, we propose a novel video face recognition algorithm, that utilizes frame selection process, followed by a deep learning architecture for feature extraction and matching as illustrated in Figure 4-3[1].

The first contribution of this research is a novel algorithm for no-reference feature-richness based frame selection that quantifies feature-richness based on entropy [178] in the wavelet domain and enables better selection of frames for recognition as compared to traditional no-reference biometric quality measures [123, 146, 147]. The second contribution is designing a novel joint feature learning framework which can be utilized to combine intermediate features computed in a deep network. Deep learning architectures generally compute a series of intermediate features from input data and utilize the final layer of feature only for representation and classification. In the proposed deep architecture, we combine the intermediate representations computed by an autoencoder using a joint representation layer. This joint representation is utilized to retain the informative features of different granularities and is used as input to a DBM which interprets and enhances this combined information to create a *feature vector* for each input face. The proposed framework models the learned features as sparse and low-rank at the same time using ℓ_1-norm and trace-norm regularizations to improve the performance of the overall deep architecture. The learnt joint representation is input to a neural network for classification. The effectiveness of the proposed algorithm is evaluated on two large publicly available benchmark databases: the YTF video [114] and PaSC video [12].

[1] A preliminary version of the proposed algorithm was published in IEEE IJCB, 2014 [62].

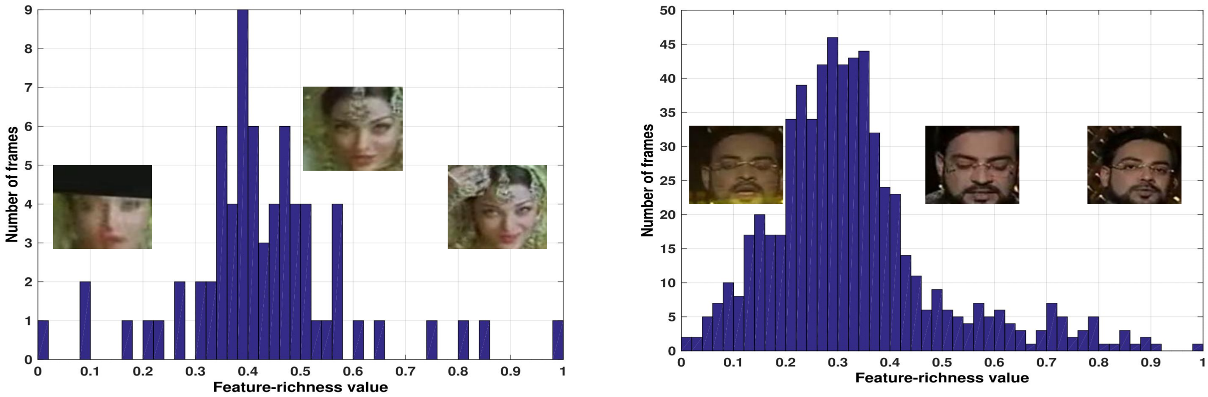


Figure 4-4: Feature-richness distributions for two different videos. Some of the most feature-rich (values close to 1) and least feature-rich frames (values close to 0) are presented for illustration. We can see that the high fidelity frames are assigned a higher feature richness score and the poor frames which showcase artifacts such as occlusion and blur are assigned a low feature-richness score. Note that the total number of frames in the two videos is different.

4.1 Proposed Face Recognition Algorithm

The proposed algorithm is divided into three steps: (i) frame selection, (ii) deep learning based feature extraction, and (iii) face verification using learnt representations. An overview of the proposed algorithm is presented in Figure 4-3.

4.1.1 Entropy based Frame Selection

Depending on the frame rate and duration, a video clip of $4-6$ seconds may contain 100-200 frames. Existing literature for video face recognition has either used all the frames, or processed some (randomly) selected frames, or have proposed algorithms for frame selection. Processing all the frames can result in inclusion of bad and redundant information. Liu *et al.* [124] proposed to partition the video into frame clusters and select the most representative frames from each cluster using PCA. Park *et al.* [204] proposed to select frames by estimating pose and motion blur information for each frame using Active Appearance Models (AAM) and selecting frames with controlled pose and minimal blur. Jillela *et al.* [97] utilized optical flow to create super-resolved frames by using short five frame sub-sequences while avoiding the sub-sequences which demonstrate high inter-frame motion.

The proposed algorithm presents a novel perspective towards frame selection by utilizing feature richness as the criteria. It is our assertion that quantifying the feature richness of an image helps in extracting the frames that have higher possibility of containing discriminatory features. In order to compute feature-richness, first the input (detected face) image I is preprocessed to a standard size and converted to grayscale. By performing face detection first and considering only the facial region, we ensure that other non-face content of the frame does not interfere with the proposed algorithm. The image is normalized using its mean and standard deviation. Thereafter, the Discrete Wavelet Transform (DWT) of the preprocessed image I is computed as follows:

$$[I_{Ap}, I_{Ho}, I_{Vr}, I_{Dg}] = DWT(I) \tag{4.1}$$

Here, I_{Ap} captures the approximation coefficients of the image, whereas $[I_{Ho}, I_{Vr}, I_{Dg}]$ contain the detail coefficients in horizontal, vertical, and diagonal sub-bands respectively. The high and low pass filters used for decomposition depend on the type of mother wavelet used. In this research, we have utilized a bi-orthogonal mother wavelet which is symmetric and efficiently encodes edge features. The detail and approximation coefficients obtained using Eq. 4.1 represent the first level DWT coefficients. Another level of DWT is applied on the approximation band, I_{Ap}, as follows:

$$[I'_{Ap}, I'_{Ho}, I'_{Vr}, I'_{Dg}] = DWT(I_{Ap}); \tag{4.2}$$

Here, I'_{Ap} and $[I'_{Ho}, I'_{Vr}, I'_{Dg}]$ represent the second level DWT approximation and detail coefficients of input image I respectively. DWT is useful to enable multi-resolution analysis of the given image. While the first level DWT presents the coefficients for the finer details of the image, the second level DWT encodes the global features while focusing less on fine details. We have observed that with images of size 80×100 and below, the third level DWT is unable to preserve sufficient edge information and is not useful for frame selection. Therefore, in this research, we consider only two levels of DWT.

For an image region, entropy signifies the variation in pixel intensity values. To quantify the feature-richness of an image, entropy [178] is computed by using both levels of DWT coefficients. The local entropy of each DWT band is computed by dividing each band into 3×3 windows. On applying the algorithm to a DWT band instead of the image, the entropy value captures the local variations in high frequency and approximation subbands contained in the image. The entropy,

$H(\kappa)$, of an image window κ is computed.

$$H(\kappa) = -\sum_{i=1}^{n} p(\kappa_i) log_2 p(\kappa_i) \tag{4.3}$$

where, n is the total number of pixel values, and $p(\kappa_i)$ is the value of the probability mass function for κ_i which represents the probability of pixel value κ_i appearing in the neighborhood. If the size of the window κ is $\mathcal{M}_\kappa \times \mathcal{N}_\kappa$ then

$$p(\kappa_i) = \frac{n_{\kappa_i}}{\mathcal{M}_\kappa \times \mathcal{N}_\kappa} \tag{4.4}$$

Here, n_{κ_i} denotes the number of pixels in the window with value κ_i. The entropy value of each window is combined to compute the feature-richness value of a band.

$$HF = \sum_{i=1}^{\omega} (|H_i|) \tag{4.5}$$

Here, HF denotes the feature-richness score of a DWT band, ω is the number of windows in the band and H_i denotes the entropy of the i^{th} window. The final score of image I, $HF(I)$, is obtained by aggregating the feature-richness values of individual bands.

$$\begin{split} HF(I) = HF(I'_{Ap}) + HF(I'_{Ho}) + HF(I'_{Vr}) + HF(I'_{Dg}) \\ + HF(I_{Ho}) + HF(I_{Vr}) + HF(I_{Dg}) \end{split} \tag{4.6}$$

Given a video $\mathcal{V}$, the feature-richness score of a frame f_i is represented as $HF(f_i)$. Since the score of each frame depends on the distribution of intensity values in a frame, it is important to normalize the scores across the frames in one video. Let m_i represent the feature-richness value corresponding to the i^{th} frame f_i, it is obtained using min-max normalization.

$$m_i = \frac{HF(f_i) - \min(\mathbf{HF})}{\max(\mathbf{HF}) - \min(\mathbf{HF})} \tag{4.7}$$

where, **HF** denotes all the feature-richness scores for the video $\mathcal{V}$ and $\min(\mathbf{HF})$ and $\max(\mathbf{HF})$ denote the minimum and maximum values in **HF**, respectively. Higher values of m signify a more feature-rich frame. Figure 4-4 shows the feature-richness distribution for two videos of different individuals from the YouTube Faces database [114] along with sample frames of high, average, and low feature-richness values. Once the score of each frame is computed, *adaptive* frame selection is performed to determine the optimum set of frames to represent a video.

Let σ_m denote the standard deviation and μ_m denote the mean pertaining to the set of feature-richness values of the video $\mathcal{V}$. In order to decide which frames are selected for verification, φ_i is computed corresponding to each frame f_i.

$$\varphi_i = \left\{ \begin{array}{ll} 1, & \text{if } m_i \geq \mu_m + \frac{\sigma_m}{2} \\ 0, & \text{otherwise} \end{array} \right\} \tag{4.8}$$

To perform adaptive frame selection, each frame with $\varphi = 1$ is selected from a given video. These frames are utilized for feature extraction using the deep learning architecture described in the next section.

4.1.2 Deep Learning Framework for Feature Extraction

Once the feature-rich frames are obtained, the next step involved feature extraction and matching. Several state-of-the-art algorithms in recent literature use convolutional neural network. In this research, we propose a SDAE and DBM based algorithm that can yield good results with limited training data while simultaneously being able to utilize more training data to further improve performance. First, we briefly present an overview of SDAE and DBM followed by the proposed architecture.

Stacked Denoising Autoencoder and Deep Boltzmann Machines

An *autoencoder* [220], [158] maps the data $\mathbf{x} \in \mathbb{R}^\alpha$ into feature (latent representation) $\mathbf{f}$ using a deterministic (encoder) function g_Θ such that,

$$\mathbf{f} = g_\Theta(\mathbf{x}) = s(\mathbf{w} \cdot \mathbf{x} + \mathbf{\Delta}) \tag{4.9}$$

where, $\Theta = \{\mathbf{w}, \mathbf{\Delta}\}$ is the parameter set, s represents the sigmoid, $\mathbf{w}$ is the $\alpha' \times \alpha$ weight matrix, and $\mathbf{\Delta}$ is the offset vector of size α'. Feature $\mathbf{f}$ can be mapped to feature vector $\hat{\mathbf{x}}$ of dimensionality

α using a decoder function $g'_{\Theta'}$ such that,

$$\hat{\mathbf{x}} = g'_{\Theta'}(\mathbf{f}) = s(\mathbf{w}' \cdot \mathbf{f} + \mathbf{\Delta}') \tag{4.10}$$

Here, $\Theta' = \{\mathbf{w}', \mathbf{\Delta}'\}$ is the decoder parameter set such that $\underset{\mathbf{w},\mathbf{w}'}{\arg\min} ||\mathbf{x} - \hat{\mathbf{x}}||_2^2$. The parameters are optimized by utilizing the unsupervised training data. Denoising autoencoder [158], a variant of autoencoder, operates on the noisy input data $\mathbf{x}_n$ and attempts to reconstruct $\hat{\mathbf{x}}$ such that $\mathbf{f} = g_\Theta(\hat{\mathbf{x}}_n) = s(\mathbf{w} \cdot \mathbf{x}_n + \Delta)$. It is observed that this variant is robust to noisy data and has good generalizability. Further, adding sparsity constraint helps in learning useful features and the cost function is updated as,

$$||\mathbf{x} - \hat{\mathbf{x}}||_2^2 + \beta \sum_j KL(\rho \parallel \hat{\rho}_j) \tag{4.11}$$

where, ρ is the sparsity parameter, $\hat{\rho}_j$ is the average activation of the j^{th} hidden unit, $KL(\rho \parallel \hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}$ is the KL-divergence, and β is the sparsity penalty term. KL divergence measures the difference between a true probability distribution and its approximation. By setting the value of ρ to a small value (such as 0.05), the number of data points for which the j^{th} unit is activated can be forced to be low, which introduces sparsity of features. Smaller values of ρ and larger values of β promote more sparse features. However, a higher value of β conversely reduces the importance of accurate reconstruction. The values of ρ and β are learnt during the training and validation stages to achieve a trade-off between reconstruction performance and learning more generalizable features. If the autoencoders are stacked in a layered manner, they are called as stacked autoencoders and form a deep learning architecture to discover "*patterns*" in the input data.

Deep Boltzmann Machine (DBM) is an undirected graphical model, a deep network architecture, with symmetrically coupled binary units [179]. It is designed by layer-wise training of RBM and stacking them together in an undirected manner. A RBM has stochastic visible and hidden variables which are connected and the energy function is defined as:

$$E(v, h; \theta) = -\sum_{i=1}^{D} \sum_{j=1}^{F} W_{ij} v_i h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{F} a_j h_j \tag{4.12}$$

Here, $\mathbf{v} \in \{0,1\}^D$ denotes the visible variables and $\mathbf{h} \in \{0,1\}^F$ denotes the hidden variables, respectively. The model parameters are denoted by $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$. W_{ij} denotes the weight of the connection between the i^{th} visible unit and j^{th} hidden unit and b_i and a_j denote the bias terms of the model. For real valued visible variables such as image pixel intensities, generally, Gaussian-Bernoulli RBMs are utilized and the energy is defined as:

$$E(v,h;\theta) = -\sum_{i=1}^{D} \frac{v_i}{\sigma_i} \sum_{j=1}^{F} W_{ij} h_j - \sum_{i=1}^{D} \frac{(v_i - b_i)^2}{2\sigma^2} - \sum_{j=1}^{F} a_j h_j \tag{4.13}$$

Here, $\mathbf{v} \in \mathbb{R}^D$ denotes the real-valued visible vector and $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}, \sigma\}$ are the model parameters. A single Gaussian-Bernoulli RBM can learn a representation of the input data. However, multiple such RBMs can be stacked in a layer-wise manner to learn increasingly complex representations of data in the form of a DBM. In this research, a three layer DBM is utilized with a greedy learning approach [54]. A three layer DBM comprised of Gaussian-Bernoulli RBMs can learn complex representations of a real-valued input vector $\mathbf{v} \in \mathbb{R}^D$ using a sequence of layers of hidden units $\mathbf{h}^{(1)}$, $\mathbf{h}^{(2)}$, and $\mathbf{h}^{(3)}$. The first layer connects the visible units to the first layer of hidden units. Thereafter, subsequent layers connect the hidden units of one layer to the hidden units of the other, causing the hidden units of a layer to act as the visible units for the next layer and so on. The energy of this DBM can be defined as:

$$\begin{aligned} E(\mathbf{v},\mathbf{h};\theta) = -&\sum_{i=1}^{D}\sum_{j=1}^{F_1} W_{ij}^{(1)} \frac{v_i}{\sigma_i} h_j^{(1)} - \sum_{j=1}^{F_1}\sum_{l=1}^{F_2} W_{jl}^{(2)} h_j^{(1)} h_l^{(2)} \\ &- \sum_{l=1}^{F_2}\sum_{m=1}^{F_3} W_{lm}^{(3)} h_l^{(2)} h_m^{(3)} - \sum_{i=1}^{D} \frac{(v_i - b_i)^2}{2\sigma^2} \\ &- \sum_{j=1}^{F_1} a_j^{(1)} h_j^{(1)} - \sum_{l=1}^{F_2} a_l^{(2)} h_l^{(2)} - \sum_{m=1}^{F_3} a_m^{(3)} h_m^{(3)} \end{aligned} \tag{4.14}$$

Here, D, F_1, F_2, F_3 are the number of units and visible and hidden layers, and $\theta = \{\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}, \mathbf{W}^{(3)}, \mathbf{b}, \mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \mathbf{a}^{(3)}, \sigma\}$ is the set of model parameters representing visible-to-hidden and hidden-to-hidden symmetric connection weights, bias terms, and the Gaussian distribution standard deviation, respectively. The probability assigned by this model to a visible vector $\mathbf{v}$ is given

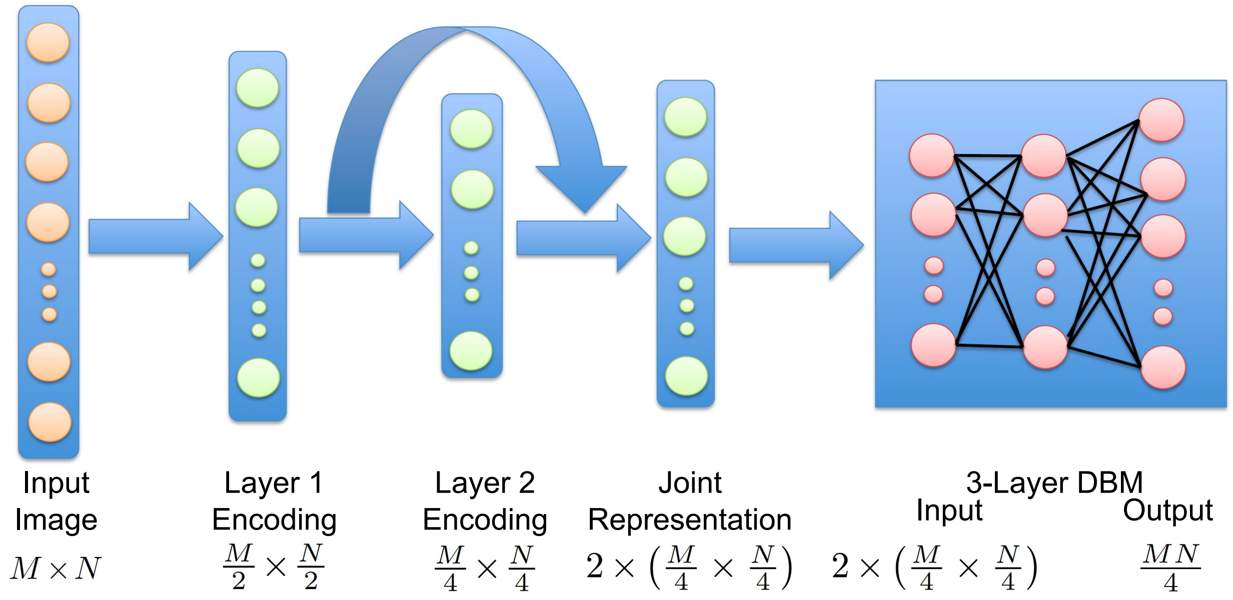

Input
Image
$M \times N$
Layer 1
Encoding
$\frac{M}{2} \times \frac{N}{2}$
Layer 2
Encoding
$\frac{M}{4} \times \frac{N}{4}$
Joint
Representation
$2 \times \left(\frac{M}{4} \times \frac{N}{4}\right)$
3-Layer DBM
Input
$2 \times \left(\frac{M}{4} \times \frac{N}{4}\right)$
Output
$\frac{MN}{4}$

Figure 4-5: Proposed deep learning architecture for facial representation: from input layer (image), two hidden layer representation are computed using SDAE encoding function. A joint representation is then obtained which combines the information from two SDAE encoding layers. Using joint representation as input, a DBM is used for computing a final feature vector.

by the Boltzmann distribution:

$$P(\mathbf{v};\theta) = \frac{1}{Z(\theta)} \sum_{\mathbf{h}} exp\left(-E\left(\mathbf{v}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{h}^{(3)}; \theta\right)\right). \tag{4.15}$$

Here, $Z(\theta)$ is the normalizing constant. If only $\mathbf{W}^{(1)}$ is considered, the derivative of the log-likelihood with respect to the model parameters is:

$$\frac{\delta log P(\mathbf{v};\theta)}{\delta \mathbf{W}^{(1)}} = \mathbb{E}_{P_{data}}[\mathbf{v}{\mathbf{h}^{(1)}}^T] - \mathbb{E}_{P_{model}}[\mathbf{v}{\mathbf{h}^{(1)}}^T] \tag{4.16}$$

Here, $\mathbb{E}_{P_{data}}[\cdot]$ denotes the expectation with respect to the data distribution and $\mathbb{E}_{P_{model}}[\cdot]$ is the expectation with respect to the distribution defined by the DBM as in Eq. (4.15). Similar derivatives are obtained for $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$, with the product $\mathbf{v}\mathbf{h}^{(1)}$ replaced by $\mathbf{h}^{(\mathbf{1})}\mathbf{h}^{(\mathbf{2})}$ and $\mathbf{h}^{(\mathbf{2})}\mathbf{h}^{(\mathbf{3})}$ respectively.

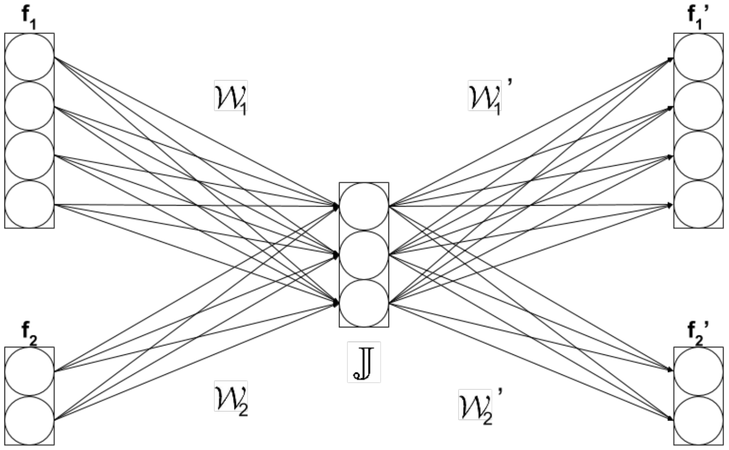


Figure 4-6: Joint learning framework: features learned from the first and second levels of autoencoder, i.e., $\mathbf{f_1}$ and $\mathbf{f_2}$ are given as input to DBM to learn the joint representation $\mathbb{J}$.

Unsupervised Joint Feature Learning

SDAE and DBM both individually learn the *useful* (intermediate) representation of input data. While the SDAE learns two layers of image-level features that can be best utilized to reconstruct the original input, in this research, we propose a joint representation layer that learns the important features from each constituent layer. This joint layer representation combines two different levels of granularities in features to obtain a better representation. Further, this joint feature is used as input to a DBM to obtain the final representation. While SDAE and joint representation are robust to noise in the input data, DBM learns the internal complex representations probabilistically. Therefore, it is our assertion that the proposed architecture should be able to produce a robust representation compared to using SDAE or DBM in isolation. Further, DBM is able to interpret the features learned by the joint representation and combine each of its components as required to obtain an enhanced higher level discriminative representation, especially after fine-tuning.

Let the size of the input data be $M \times N$; in the proposed architecture, each layer of SDAE is one-fourth the size of its previous layer. Layer-by-layer greedy approach [53] with stochastic gradient descent is utilized to train the SDAE followed by fine-tuning with back-propagation method. Intermediate representations obtained using the 2-hidden layer SDAE are further combined to obtain a joint representation as illustrated in Figure 4-5. The two layers of size $\frac{M}{2} \times \frac{N}{2}$ and $\frac{M}{4} \times \frac{N}{4}$ are utilized as input and one joint layer of size $2 \times \left(\frac{M}{4} \times \frac{N}{4}\right)$ is learned. Let $\mathbf{f_1}$ be the representation learned by the first layer of SDAE and $\mathbf{f_2}$ be the feature learned by the second layer of SDAE, the joint representation $\mathbb{J}$ can be learned using Eq. (4.17).

$$\mathbb{J} = \mathcal{G}(\mathbf{f_1}, \mathbf{f_2}) \tag{4.17}$$

Here, $\mathcal{G}$ is the joint learning function to obtain $\mathbb{J}$. In this research, using encoder-decoder approach, we defined the cost function associated with Eq. (4.16) as:

$$\underset{\Phi}{\operatorname{argmin}}(\| \mathbf{f_1} - \mathbf{f_1'} \|_2^2 + \| \mathbf{f_2} - \mathbf{f_2'} \|_2^2 + \mathcal{R}) \tag{4.18}$$

where, Φ represents the set of all the variables to be learnt and $\mathcal{R}$ is a regularizer. For ease of explanation, we first present the formulation with linear activation. Eq. (4.17) can be written as,

$$\mathbb{J} = \mathcal{W}_1\mathbf{f_1} + \mathcal{W}_2\mathbf{f_2} \tag{4.19}$$

Using Eq. (4.18), the associated cost can be written as,

$$\begin{multline}\underset{\Phi}{\operatorname{argmin}}(\| \mathbf{f_1} - \mathcal{W}_1'\mathcal{W}_1\mathbf{f_1} - \mathcal{W}_1'\mathcal{W}_2\mathbf{f_2} \|_2^2 + \\ \| \mathbf{f_2} - \mathcal{W}_2'\mathcal{W}_2\mathbf{f_2} - \mathcal{W}_2'\mathcal{W}_1\mathbf{f_1} \|_2^2 + \mathcal{R}) \quad (4.20)\end{multline}$$

As shown in Figure 4-6, this approach learns the weights $\Phi = \{\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_1', \mathcal{W}_2'\}$ to obtain the joint representation $\mathbb{J}$. In a similar fashion, non-linear cost function can be written as (for simplicity, bias terms are omitted),

$$\begin{multline}\underset{\Phi}{\operatorname{argmin}}(\| \mathbf{f_1} - s(\mathcal{W}_1'[s(\mathcal{W}_1\mathbf{f_1})]) - s(\mathcal{W}_1'[s(\mathcal{W}_2\mathbf{f_2})]) \|_2^2 + \\ \| \mathbf{f_2} - s(\mathcal{W}_2'[s(\mathcal{W}_2\mathbf{f_2})]) - s(\mathcal{W}_2'[s(\mathcal{W}_1\mathbf{f_1})]) \|_2^2 + \mathcal{R}) \quad (4.21)\end{multline}$$

Adding ℓ_2-norm regularization term on $\mathcal{W}_1, \mathcal{W}_2$ and *dropout* [190] on the joint representation network, Eq. (4.21) can be written as,

$$\underset{\Phi}{\operatorname{argmin}} \big(\| \mathbf{f_1} - s(\mathcal{W}'_{\mathbf{1}}[s(\mathcal{W}_{\mathbf{1}}\mathbf{f_1})]) - s(\mathcal{W}'_{\mathbf{1}}[s(\mathcal{W}_{\mathbf{2}}\mathbf{f_2})]) \|_2^2 + \\ \| \mathbf{f_2} - s(\mathcal{W}'_{\mathbf{2}}[s(\mathcal{W}_{\mathbf{2}}\mathbf{f_2})]) - s(\mathcal{W}'_{\mathbf{2}}[s(\mathcal{W}_{\mathbf{1}}\mathbf{f_1})]) \|_2^2 + \\ (\lambda_1 \| \mathcal{W}_{\mathbf{1}} \|_2^2 + \lambda_2 \| \mathcal{W}_{\mathbf{2}} \|_2^2)\big)_{dropout} \tag{4.22}$$

The joint representation combines *abstract* and *low-level* features obtained from SDAE encoding layers and is used as input to a three hidden layer DBM, i.e. $\mathbb{J}$ acts as the visible vector. Similar to Eq. (4.14), the energy of this DBM is represented as:

$$E(\mathbb{J}, \mathbf{h}; \theta) = -\sum_{i=1}^{D}\sum_{j=1}^{F_1} W_{ij}^{(1)} \frac{\mathbb{J}_i}{\sigma_i} h_j^{(1)} - \sum_{j=1}^{F_1}\sum_{l=1}^{F_2} W_{jl}^{(2)} h_j^{(1)} h_l^{(2)} \\ -\sum_{l=1}^{F_2}\sum_{m=1}^{F_3} W_{lm}^{(3)} h_l^{(2)} h_m^{(3)} - \sum_{i=1}^{D} \frac{(\mathbb{J}_i - b_i)^2}{2\sigma^2} \\ -\sum_{j=1}^{F_1} a_j^{(1)} h_j^{(1)} - \sum_{l=1}^{F_2} a_l^{(2)} h_l^{(2)} - \sum_{m=1}^{F_3} a_m^{(3)} h_m^{(3)} \tag{4.23}$$

Inspired from [163, 181], we believe that the learned weight matrix can be modeled as sparse and low rank [223] at the same time and therefore, a regularization approach incorporating both of these can improve feature learning. Hence, we extend the loss function of DBM (RBM) by introducing trace-norm regularization technique.

Let $\mathcal{L}$ be the loss function of RBM (DBM) with the energy function defined in Eq. (4.23). Along with ℓ_1-norm, trace-norm is added to the loss function as follows:

$$\mathcal{L}_{new} = \mathcal{L} + \mathcal{A} \| W \|_1 + \mathcal{B} \| W \|_\tau \tag{4.24}$$

where $\| \cdot \|_1$ is the ℓ_1-norm, and $\| \cdot \|_\tau$ is the trace-norm, and $\mathcal{A}$, $\mathcal{B}$ are the regularization parameters which control sparsity and low-rankness. In general, elastic net regularization ($\| \cdot \|_1 + \| \cdot \|_2$) [245] may be used; however in this formulation, we propose to utilize trace-norm in conjunction with ℓ_1-norm for learning representation in RBM (DBM). While ℓ_1-norm induces sparsity in the weight matrix, trace-norm induces features to have low-rankness. The weight matrix learned by the updated loss function has the benefits of both the regularizations and as shown in experimental results, improves the overall verification performance.

The size of the first two layers of the DBM is set to $2 \times \left(\frac{M}{4} \times \frac{N}{4}\right)$ and the final layer is set to $\frac{MN}{4}$. A pre-training approach [54] combined with generative fine-tuning [79] is followed to train the DBM. The final hidden layer provides a complex representation of the input which can be utilized for classification.

4.1.3 Face Verification using Feature Richness and Deep Learning based Representation

As shown in Figure 4-3, the proposed framework utilizes the frame selection, feature extraction, and classification architecture for video based face recognition. During training, the stack of SDAE joint representation and DBM is utilized for facial representation. Let $I_{gallery}$ and I_{probe} be the two detected, preprocessed and geometrically normalized face images to be matched. These images are resized to $M \times N$ (in our experiments, it is 80×100) and converted into vector form. The trained architecture is used to extract the features from $I_{gallery}$ and I_{probe}, respectively. According to the previous discussion, the input to the feature extraction module is the MN size image vector and the output is a vector of length $\left(\frac{MN}{4}\right)$. Features are extracted for each selected frame in a video and given as input to a five layer neural network (one input layer - 3 hidden layers - one output layer) for classification (verification). The neural network classifier is trained to verify a pair of input images (frames) input as a concatenated feature vector of size $\frac{MN}{2}$, using all the frames in the training videos. The output of the network is a scalar match score.

During testing, the most feature-rich frames are selected from each of the gallery and probe videos, and matched using the proposed feature extraction and matching algorithm. The output of neural network (classifier) is undecimated and match scores are computed. The videos to be matched may have significant variations in quality and feature-richness. It has been shown in literature that if the images are of very different quality, then the matching performance may deteriorate [17]. Therefore, we perform a post-processing step to select frame-pairs with similar feature-richness and discard the remaining pairs. Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be the two videos to be matched,

Table 4.2: Details of the YTF and PaSC databases.

Database	No. of		Avg. no. of	
	Subjects	Videos	Videos per subject	Frames per video
YouTube Faces	1595	3425	2.15	181.3
PaSC (Handheld)	265	1401	4 to 7	234.8
PaSC (Control)	265	1401	4 to 7	239.0

a pair-wise feature-richness value is computed for each possible frame-pair using the algorithm explained in Section 4.1.1.

$$\left[m_{1,1}m_{1,2}; m_{2,1}m_{2,2}; ..., m_{i,1}m_{j,2}; ..., m_{\mathcal{N}_{1,1}}m_{\mathcal{N}_{2,2}}\right] \tag{4.25}$$

$m_{i,1}m_{j,2}$ denotes the product of feature-richness value associated with the pair formed by the i^{th} frame from $\mathcal{V}_1$ and the j^{th} frame from $\mathcal{V}_2$. $\mathcal{N}_1$ and $\mathcal{N}_2$ denote the total number of selected frames from $\mathcal{V}_1$ and $\mathcal{V}_2$ respectively. Let $\sigma^{'}_{m}$ be the standard deviation and $\mu^{'}_{m}$ be the mean pertaining to the set of the pair-wise feature-richness values for all pairs possible between $\mathcal{V}_1$ and $\mathcal{V}_2$. To finally select the pairs for decision making, following equation is utilized:

$$\Upsilon_{i,j} = \begin{Bmatrix} 1, & \text{if } m_{i,1}m_{j,2} \geq \mu^{'}_{m} + \frac{\sigma^{'}_{m}}{2} \\ 0, & \text{otherwise} \end{Bmatrix} \tag{4.26}$$

If the combined score of a pair $f_{i,1}f_{j,2}$ is more than the threshold, i.e., if $\Upsilon_{i,j} = 1$, then this pair is considered for computing the match score. While pairs with $\Upsilon_{i,j} < 1$ are not considered for verification, other selected frame-pairs are weighted according to the joint feature-richness value. For frame-pair $f_{i,1}f_{j,2}$, this weight is computed as $\Upsilon_{i,j}m_{i,1}m_{j,2}$. A pair where both participating frames are highly feature-rich is assigned a higher weight compared to other combinations. Here, facial coordinates obtained during face detection are used to ensure that frontal-only and semi-profile images are not matched with profile faces (i.e, when pose variations are very large). The final match score is computed in the form of a weighted sum of scores obtained from each participating frame-pair. The undecimated/unthresholded network (classifier) output of these pairs are combined using weighted sum rule [175] and a verification threshold is applied to provide the final decision of accept or reject (same or not same) at a fixed false accept rate.

4.2 Results and Analysis

In order to evaluate the efficacy of the proposed algorithm, face verification[2] experiments are performed on two popular video benchmark databases: YouTube Faces [114] and the Point and Shoot Challenge [12]. Three different experiments are performed to demonstrate the efficacy of the proposed algorithm.

- compare the performance of state-of-the-art results reported on these databases with the proposed algorithm,
- evaluate the effectiveness of individual components of the proposed algorithm, and
- evaluate the generalization capability by evaluating the performance with cross database training and testing sets.

4.2.1 Database and Experimental Protocol

The YouTube Faces database contains 3,425 videos downloaded from YouTube belonging to 1,595 individuals. The PaSC database contains 1,401 handheld and 1,401 control (high resolution) videos pertaining to 265 individuals. Videos in the PaSC database capture individuals in various indoor and outdoor locations while performing a predefined activity. The details of both the databases are summarized in Table 4.2. Both YouTube Faces database [114] and PaSC database [12] have predefined experimental protocols. For the YouTube faces database, we have followed the restricted protocol which consists of 10 splits, each containing 250 genuine and 250 impostor pairs. No information outside of these splits is used during any stage of the evaluation. The results are reported with 10 fold cross validation, 9 splits are used for training and one split for testing.

The PaSC database contains videos from a handheld camera of low resolution and a control camera of high resolution. The handheld-to-handheld experiment evaluates the accuracy of an algorithm when matching videos of low resolution, whereas the control-to-control evaluates the accuracy for high resolution videos. The experiments are performed for both handheld-to-handheld and control-to-control protocols of video face recognition. Training is performed on a separate set

[2] In biometrics, recognition has two components: verification (1:1 matching) and identification (1:N) matching. In this research, we have interchangeably used verification and recognition to report 1:1 matching performance.

of training videos provided with the database and the signature sets already provided with the PaSC database are used to select the pairs for testing.

For both databases, the training data is divided into two parts: first part is utilized as unlabeled data for training the proposed joint representation model and second part is used for supervised training. Data augmentation with different image processing operations such as mirror/flip, color/grayscale, and jittering, are also applied to increase the training database size. After training, the proposed algorithm is evaluated on the testing data. The metadata and annotations provided with each database are used to perform face detection and pose detection (to determine pose) as applicable. Receiver Operating Characteristic (ROC) curves are computed for each experiment and the verification accuracies are reported at multiple false accept rates.

4.2.2 Experimental Results

Results on YouTube Faces Database

ROC curves of existing algorithms and the proposed face recognition algorithm on the YouTube faces database are shown in Figure 4-7. It is evident that the proposed algorithm not only achieves high accuracy at low FARs, but also achieves state-of-the-art performance of 0.93 Genuine Accept Rate (GAR) at equal error rate, without outside training data. We next analyzed selected top-performing algorithms to understand their performance at 0.01 FAR which is more pragmatic with respect to real world scenarios. Since the absolute performance of algorithms has improved overall, it is easier to compare their performance at 0.01 FAR as opposed to EER to determine which algorithm has better performance. This can be seen in Figure 4-2 as well where the difference between the top two algorithms at EER is just 2% and they seem to be similarly accurate but the delta increases to 25% when performance at 0.01 FAR is considered and it is easier to compare them. As shown in Figure 4-8, the proposed algorithm substantially outperforms these algorithms at lower FARs. At 0.01 FAR, the proposed algorithm achieves GAR of 0.79 whereas, the next best GAR is 0.54 by DeepFace[3]. It is our assertion that selection of feature-rich frames and the proposed joint representation architecture helps to yield state-of-the-art face verification performance.

[3] Since the ROC curve of FaceNet is not available, the results of FaceNet at different FARs could not be reported.

Results on PaSC Database

As explained in Section 3.1, Point and Shoot Challenge database has two protocols: handheld and control. Table 4.3 summarizes the results of the proposed algorithm along with existing results reported on both the protocols. Beveridge *et al.* [12] reported the performance of PittPatt and Local Region Principal Component Analysis (LRPCA) on both handheld and control subsets. The results show that at 0.01 FAR, the GAR of the proposed algorithm is more than twice of PittPatt. At 0.01 FAR, the proposed algorithm yields 0.93 and 0.96 GAR on the handheld and control subsets, respectively. Beveridge et al. [11], [13] have reported the results of the PaSC Video Face and Person Recognition Competitions. Table 4.3 shows the genuine accept rates of the algorithms reported in the competitions along with the results of the proposed algorithm. These results show that the

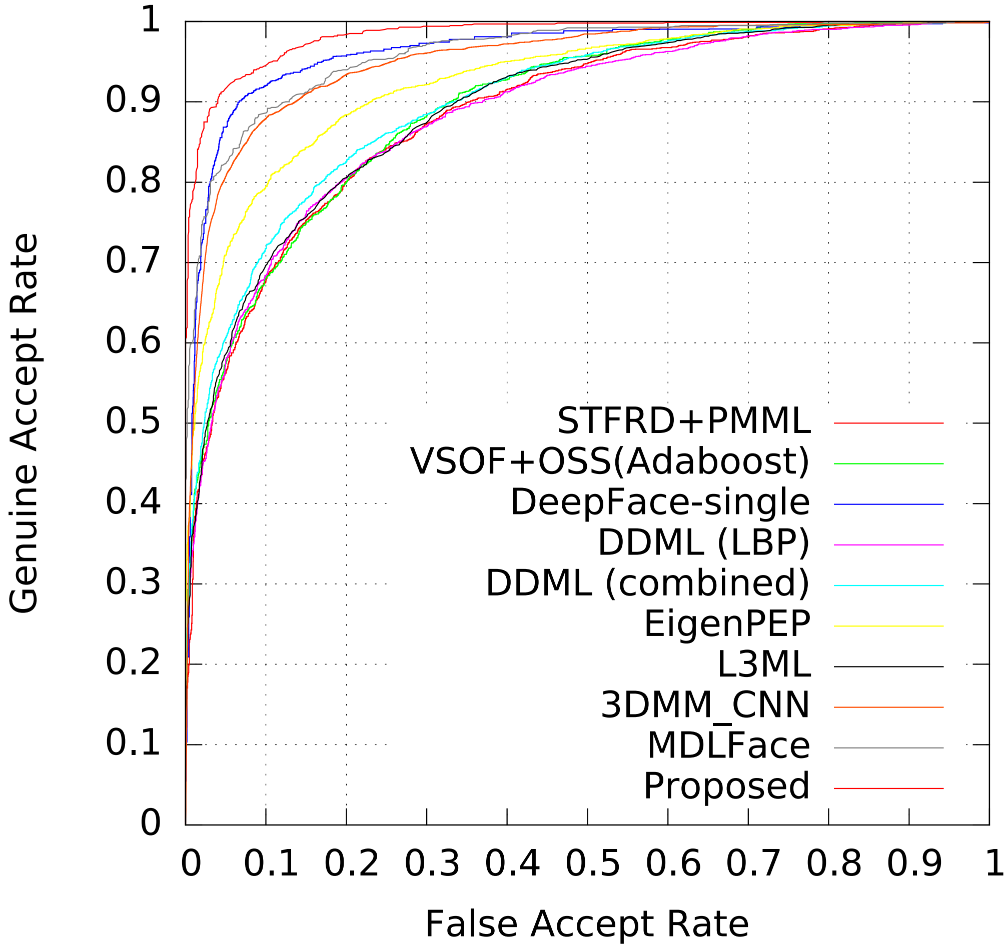


Figure 4-7: ROC curves comparing the verification performance of the proposed algorithm with existing results reported on the YTF database webpage.

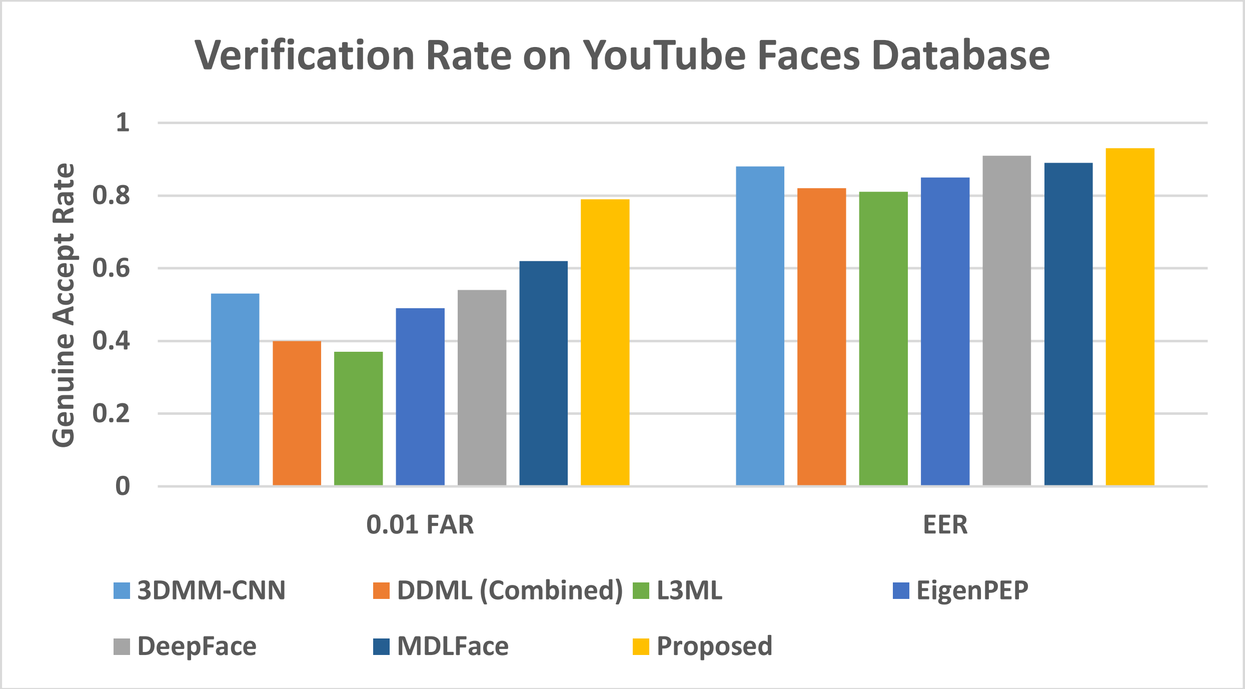


Figure 4-8: Summarizing the verification performance of the proposed algorithm and state-of-the-art algorithms on the YouTube Faces database.

proposed algorithm yields at least 34% higher verification accuracy than existing algorithms that have not utilized external data for training.

Impact of Frame Selection

Frame selection is an integral component of the proposed algorithm. The algorithm selects feature-rich frames from the given video and utilizes them for video to video matching. To evaluate the effectiveness of the proposed frame selection algorithm, multiple experiments are performed, including comparison with standard image quality measures.

Ideally, if the frames are selected optimally, then they should yield the best verification performance. To evaluate this, we have compared the verification performance of the proposed feature-rich frames with only frontal frames and when frames are selected randomly. Figure 4-9 shows sample frames from the PaSC database. It illustrates randomly selected frames, frontal frames, most feature-rich frames and the least feature-rich frames as well. It can be observed that the most feature-rich frames are distinct in nature and of good quality whereas, the least feature-rich frames

Table 4.3: Verification rates on the PaSC database. Results of existing algorithms are reported from respective references.

Algorithm	GAR at 0.01 FAR	
	Handheld	Control
ISV-GMM [13]	0.05	-
LBP-SIFT-WPCA-SILD [13]	0.09	-
PLDA-WPCA-LLR [13]	0.19	-
Eigen-PEP [13]	0.26	-
LRPCA Baseline [12]	0.08	0.10
PittPatt Baseline [12]	0.38	0.49
Surrey [11]	0.13	0.20
SIT [11]	0.31	0.35
Uni-Lj [11]	0.33	0.39
UTS [11]	0.38	0.48
CAS [11]	0.59	0.58
MDLFace [62]	0.89	0.94
Proposed	**0.93**	**0.96**

computed using the proposed frame selection algorithm do not contain very distinguishing information and are of poor quality. It is also interesting to note that the most feature-rich frames are not necessarily the frontal frames. The experiments are performed with both YouTube and PaSC databases and the results are presented in Figure 4-10. It is evident that selecting the most feature-rich frames provides the best performance across all three protocols. Correlating these images with the accuracies re-emphasizes our hypothesis that frontal frames are not always optimal and hence do not necessarily provide the best verification results.

We also compare the performance of the proposed frame selection approach with frame selection based on no-reference image quality metrics namely BRISQUE [146], NIQE [147], and Spatial-Spectral Entropy-based Quality (SSEQ) [123]. The source codes provided by the respective authors have been utilized for each of these approaches. Similar to the proposed approach, frames are selected based on the quality measure and used in the proposed framework. We have also evaluated the performance of our preliminary frame selection approach [62] and the verification results obtained with each of the frame selection algorithms and the proposed face recognition algorithm are presented in Table 4.4. We observe that using any of the existing quality assessment algorithms results in a noticeable decline in the verification performance. On the YouTube faces

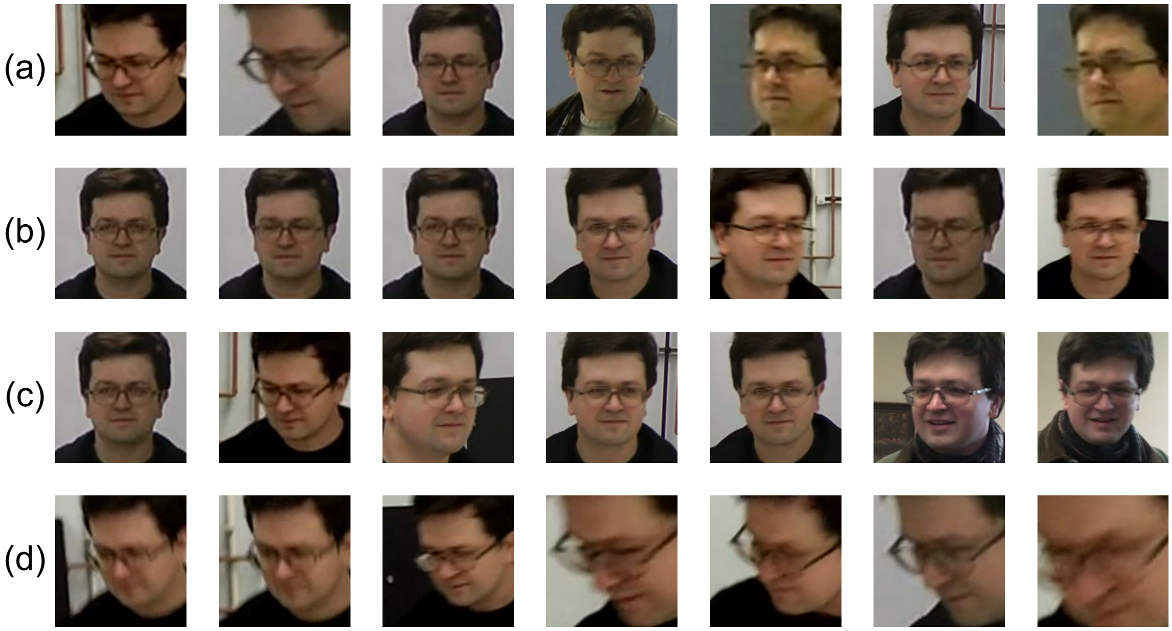
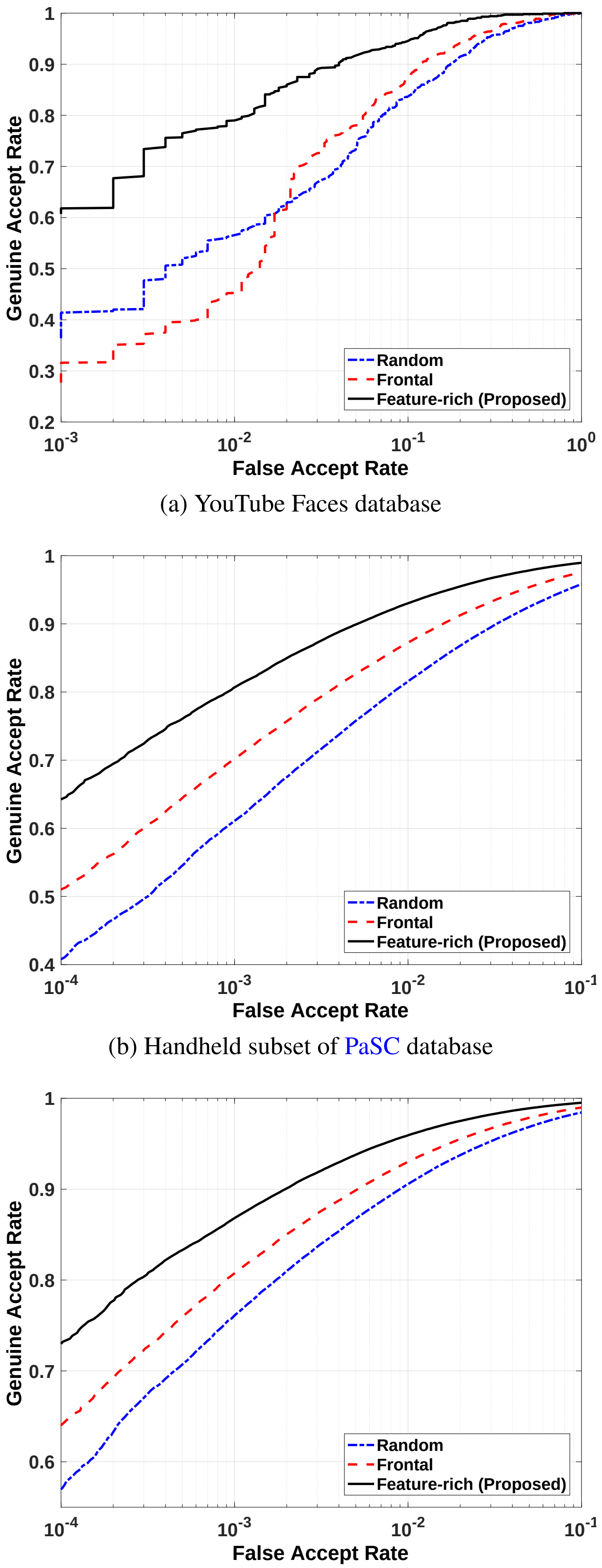


Figure 4-9: Sample frames from the PaSC database: (a) random frames, (b) frontal frames, (c) most feature-rich frames, and (d) least feature-rich frames.

database, the performance varies from 0.62 to 0.79 GAR, whereas on the handheld subset of the PaSC database the performance varies from 0.82 to 0.93 GAR by only changing the frame selection approach. The proposed feature-richness based frame selection approach consistently outperforms the quality based measures on all the protocols of both the databases. This experiment suggests that high image quality may not represent high feature richness and can affect the overall verification performance. This is consistent with existing observations in *biometrics quality* literature [17]. We further analyze the performance of the proposed algorithm with fixed number of frames i.e., without adaptive approach, as well as without using any frame selection. As shown in Table 4.4, with all frames, top-25 and top-50 feature-rich frames, the verification accuracies are relatively lower. This shows the usefulness of the “adaptive" nature of the proposed algorithm. These experiments also validate our hypothesis that not all frames are useful for video face recognition.

Analysis of Deep Learning Architecture

Individual components of the proposed deep learning framework are experimentally evaluated to determine the efficacy of the algorithms. In this experiment, only one component is changed and



(a) YouTube Faces database

(b) Handheld subset of PaSC database

(c) Control subset of PaSC database

Figure 4-10: ROC curves comparing the verification performance of the proposed algorithm with frame selection approaches on the two databases.

Table 4.4: Comparing the results of the proposed frame selection algorithm with existing image quality assessment algorithms and random frame selection.

Frame Selection	Algorithm	GAR at 0.01 FAR		
		YTF	PaSC Handheld	PaSC Control
All		0.74	0.89	0.92
Image Quality	BRISQUE [146]	0.62	0.82	0.84
	NIQE [147]	0.62	0.83	0.82
	SSEQ [123]	0.62	0.82	0.82
Memorability	MDLFace [62]	0.69	0.89	0.94
Proposed Feature Richness	25	0.75	0.91	0.94
	50	0.77	0.91	0.93
	Adaptive	**0.79**	**0.93**	**0.96**

the remaining components of the proposed framework are left unchanged and only the feature extractor module is varied across different experiments. These components include: (a) single layer denoising autoencoder, (b) two layer SDAE, (c) DBM, and (d) SDAE+DBM without the proposed joint representation layer.

Table 4.5 summarizes the GAR at 0.01 FAR for each of these components on both YouTube and PaSC databases (using feature-rich frames). From the results, it is evident that both SDAE and DBM are required in the proposed architecture to extract meaningful representation for face recognition. Using only DBM provides better performance than only using a 2-layer SDAE. However, neither DBM nor SDAE is able to achieve even 50% verification accuracies individually. A significant improvement is observed when SDAE and DBM are stacked sequentially. The proposed joint representation further improves the performance of the architecture, resulting in an improvement of up to 0.18 in GAR for the YouTube faces database. As mentioned previously, the joint representation combines different layers of feature granularity and from the results, it is evident that it is able to further improve upon the features learned by the deep architecture. This observation strengthens the requirement for the additional layer of learning after SDAE before the features are utilized by DBM.

An additional experiment is performed to evaluate the efficacy of the addition of trace-norm regularization. For this experiment, ℓ_2-norm, ℓ_1-norm, elastic net ($\ell_1 + \ell_2$ norm), trace-norm (ℓ_τ)

Table 4.5: Analyzing the performance of individual components of the proposed algorithm for face recognition.

Modified Architecture	GAR at 0.01 FAR		
	YouTube	PaSC	
		Handheld	Control
1 Layer DAE only	0.21	0.09	0.12
2 Layer SDAE only	0.39	0.28	0.39
DBM only	0.41	0.48	0.49
SDAE+DBM	0.61	0.87	0.93
Proposed: SDAE+DBM with joint representation	**0.79**	**0.93**	**0.96**

Table 4.6: GAR for cross database experiments at 0.01 FAR.

Training Set	Testing Set		
	YTF	PaSC-Handheld	PaSC-Control
YTF	0.79	0.72	0.78
PaSC	0.43	0.93	0.96
PaSC + YTF	**0.83**	**0.96**	**0.97**

only, and $(\ell_1 + \ell_\tau)$ are evaluated in the proposed framework (as shown in Eq. 4.24). For these regularizers, we observe that $(\ell_1 + \ell_\tau)$ yields the best results followed by elastic net. Incorporating single norms i.e., ℓ_1-norm and ℓ_2-norm only, yield almost similar performance and are 1-2% (at 1% FAR) less than $(\ell_1 + \ell_\tau)$ regularization.

The number of parameters in a deep neural network is determined by the weights and bias of each layer. The proposed algorithm involves a total of 22.5 million parameters whereas, other deep architectures such as Deepface [222] contain many more parameters (e.g. 120 million for Deepface). We observe that even with a relatively small number of parameters, the proposed algorithm achieves higher performance than Deepface. While architectures proposed in [160] and [183] perform better on the YouTube database than the proposed algorithm, both involve a much higher number of parameters and have utilized large amounts (2.6 million and 200 million images respectively) of training data (the results are reported on the unrestricted setting of YouTube). It is to be noted that for these experiments, the proposed algorithm is not trained with external training data.

Table 4.7: Comparing the verification accuracy of recent CNN based methods with the proposed algorithm.

Algorithm	External Data	Layers	YTF (at EER)	PaSC (at 1% FAR)	
				Control	Handheld
Trunk-Branch Ensemble CNNs with Batch Normalization [42]$^{\#}$	2.68 Million$^{\$}$	18 + 11 + 11*	94.9	98.0	97.0
VGG Face [160]$^{+}$	2.62 Million	21	**97.4**	91.3	87.0
GoogLeNet [196] features with aggregation [227]	3 Million	22	95.5	-	-
CNN-3DMM Estimation [201]	0.49 Million	101	88.8	-	-
Proposed SDAE-DBM Joint Representation	No	9	93.4	95.9	93.1
	YTF + PaSC	9	95.0	96.6	96.1
	2.48 Million	9	95.4	**98.1**	**97.2**

$^{\#}$*Results on YTF are obtained from [42], results on PaSC are obtained from [182].* $^{\$}$*2.68 million images obtained by augmenting 0.49M original images from the CASIA-WebFace database [229] using horizontal flipping and image jittering as explained in [42].* **The method uses a primary network with 18 layers and two secondary networks with 11 layers each.* $^{+}$*PaSC results are obtained from [42].*

Cross Database Experiments

The generalizability of an algorithm can be evaluated in situations where the training and testing data belong to different databases, i.e, cross-database experiments. To evaluate the effectiveness of the proposed algorithm in cross database scenarios, we have performed three different experiments:

- Training and testing databases belong to the same database. For instance, training with YouTube faces train set and testing with YouTube faces test set.

- Training and testing databases belong to different database. For instance, training with YouTube faces train set and testing with PaSC test set.

- Training database is from multiple databases whereas, the testing is performed with a single database. For instance, training with both YouTube faces and PaSC train sets and testing on YouTube faces test set.

The results of all three experiments are presented in Table 4.6. On training with the YouTube Faces database and testing with the PaSC database, the proposed algorithm yields 0.72 GAR at 0.01 FAR which is considerably better than the results of many existing algorithms. On the other hand, the performance on the Youtube faces database suffers heavily when training data is taken only from the PaSC database. This may be due to the fact that the overall quality of faces in the Youtube video faces database is lower than the training set of the PaSC challenge database. Since the representation module has not seen low quality frames and noisy faces during training, it is unable to perform well on the YouTube database. On combining the training set from both the databases, i.e. PaSC + YTF training, the accuracies of both testing cases are improved. This is a well understood phenomena in deep learning - more training data is useful in improved representation and thereby achieving higher accuracies.

4.2.3 Comparison with Recent CNN based Algorithms

We next compare the performance of the proposed algorithm with some recently proposed CNN based algorithms on the benchmark protocols of the YTF and PaSC face databases. As shown in Table 4.1, convolutional neural networks have demonstrated state-of-the-art results in deep learning based video face recognition; however, they generally use external data for training. Therefore, we have reported the results of the proposed algorithm in three settings: (i) without any external training data, (ii) using YTF and PaSC for training (as discussed in Section III-B), and (iii) using external training data of 2.48 million (with augmentation).

Table 4.7 summarizes the results of the proposed and existing algorithms. Results of existing algorithms are reported directly from the associated publications, and the results for [160] are taken from [42]. Since we have not manually pruned the PaSC database for falsely detected faces, we report the corresponding performance values for [42]. We observe that even without utilizing any external data, the proposed algorithm is able to achieve comparable accuracies. Using large training data, the accuracy improves and with 2.48 million training data, the verification rate is higher compared to existing algorithms. In terms of computational requirements, on a 32 core server with Tesla K80 Graphics Processing Unit (GPU) with 512 GB RAM, the proposed algorithm requires approximately 29 hours to train with external data. Once the model is trained, it requires about 2 seconds to match two videos.

On analyzing the architectures, we observed that in order to optimize the network for a given problem, a deep CNN architecture requires a large number of layers, which results in a large number of parameters to optimize. This requires large number of training data so that all the parameters of the network can be estimated without overfitting. The proposed algorithm achieves comparable performance with a network of lesser depth (9 layers as compared to 22 layers in [227]) with relatively less training data. We also assert that the proposed architecture can be applied to solve other challenging problems where relatively less labeled data is available such as newborn face recognition [16].

4.3 Summary

Verifying identities in videos has several applications in social media, surveillance, and law enforcement. Existing approaches have achieved high verification accuracies at equal error rate; however, achieving high performance at low false accept rate is still an arduous research challenge. In this research, a novel video face verification algorithm is proposed which utilizes frame selection and deep learning based feature representation. The proposed algorithm starts with adaptively selecting feature-rich frames from input videos using wavelet decomposition and entropy. The proposed deep learning architecture which combines SDAE joint representation with DBM is used to extract features from the selected frames. The extracted representations from two videos are matched using a feed forward neural network. The results are demonstrated on the challenging PaSC and YTF databases. The comparison with state-of-the-art results on both the databases show that the proposed algorithm provides the best results on both the databases at low false accept rate, even with limited training data. Apart from the benchmark protocols of both the databases, several additional experiments have been performed to show the effectiveness of the proposed contributions: (i) joint feature learning in an autoencoder, (ii) sparse and low rank regularization in DBM, and (iii) combination of SDAE and DBM in the proposed architecture.

Chapter 5

Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks

Deep *learning* paradigm has seen significant proliferation in face recognition due to the convenience of obtaining large training data, availability of inexpensive computing power and memory, and utilization of cameras at multiple places. Several algorithms such as DeepFace [222], DeepID [192], FaceNet [183], and Liu *et al.* [122] are successful examples of the coalesce of deep learning and face recognition. However, it is also known that machine learning algorithms are susceptible to *adversaries* which can cause the classifier to yield incorrect results. One class of such adversaries is spoofing based adversaries which use precaptured face images or videos to pass an attacker off as a different individual [4, 140]. This research focuses on exploring and handling the other class of attack where the accuracy of the targeted system is compromised by using adversarial input images. Most of the time these adversaries are unintentional and are in the form of outliers. Recently, it has been shown that *fooling images* can be generated in such a manner where humans can correctly classify the images but deep learning algorithms misclassify them [60], [151]. Such images can be generated via evolutionary algorithms [151] or adversarial sample crafting using the fast gradient sign method [60]. Sharif et al. [185] explored threat models by creating *perturbed eye-glasses* to fool face recognition algorithms. An adversarial attack on face recognition is not acceptable as face biometric gets used in many high security applications such as passports, visa,

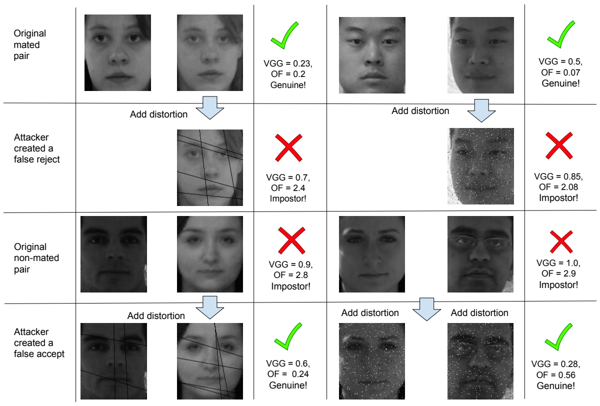

Original mated pair
VGG = 0.23, OF = 0.2 Genuine!
VGG = 0.5, OF = 0.07 Genuine!
Add distortion
Add distortion
Attacker created a false reject
VGG = 0.7, OF = 2.4 Impostor!
VGG = 0.85, OF = 2.08 Impostor!
Original non-mated pair
VGG = 0.9, OF = 2.8 Impostor!
VGG = 1.0, OF = 2.9 Impostor!
Add distortion
Add distortion
Add distortion
Attacker created a false accept
VGG = 0.6, OF = 0.24 Genuine!
VGG = 0.28, OF = 0.56 Genuine!

Figure 5-1: We show that deep learning based OpenFace and VGG can be deceived even by image processing operations that mimic real world distortions.

and other law enforcement documents.

It is our assertion that it is not required to attack the system with sophisticated learning based attacks; attacks such as adding random noise or horizontal and vertical black grid lines in the face image cause reduction in face verification accuracies. Samples shown in Figure 5-1 show a glimpse of the effect of image processing operations on two state-of-the-art deep learning based face recognition algorithms. To the best of our knowledge, this is the first reported research on finding singularities in deep learning based face recognition engines along with detection and mitigation of such attacks. We believe that being able to not only automatically detect but also correct adversarial samples at runtime is a crucial ability for a deep network that is deployed for real world applications. With this research, we aim to present a new perspective on potential attacks as well as a different methodology to limit their performance impact beyond simply including adversarial samples in the training data.

The objective of this research is three-fold: (i) we demonstrate that the performance of deep learning based face recognition algorithms can be significantly affected due to adversarial attacks - both image processing based adversarial attacks and adversarial samples generated in context to the recognition architecture. (ii) The first key step in taking countermeasures against such adversarial attacks is to be able to reliably determine which images contain such distortions. We propose and evaluate a methodology for automatic detection of such attacks using the response from hidden layers of the DNN. (iii) Once identified, the distorted images may be rejected for further processing or rectified using appropriate preprocessing techniques to prevent degradation in performance. To

Table 5.1: Literature review of adversarial attack generation and detection algorithms.

Adversary	Authors	Description
Generation	[197]	L-BFGS: $L(x+\rho,l)+\lambda\vert\vert\rho\vert\vert^2\ s.t.\ x_i+\rho_i \in [b_{min}, b_{max}]$
	[60]	FGSM: $x_0+\epsilon*(\nabla_x L(x_0,l_0)$
	[113]	I-FGSM: $x_{k+1}=x_k+\epsilon*(\nabla_x L(x_0,l_0)$
	[159]	Saliency Map: l_0 distance optmization
	[149]	DeepFool: $for\ each\ class, l \neq l_0, minimize\ d(l,l_0)$
	[22]	C & W: l_p distance metric optimization
	[148]	Universal: Distribution based perturbation
	[173]	Blackbox: Uniform, Gausaaian, Salt and Pepper, Gaussian Blur, Contrast
Detection	[70]	Statistical test for attack and genuine data distribution
	[59, 145]	Neural network based classification
	[51]	Randomized network using Dropout at both training and testing
	[14]	PCA based dimensionality reduction algorithm
	[120]	Quantization and smoothing based image processing
	[128]	Quantize ReLU output for discrete code + RBF SVM
	[36]	JPEG compression to reduce the effect of adversary

address this challenge without increasing the failure of process rate (by rejecting the samples), the third contribution of this research involves a novel technique of selective dropout in the DNN to mitigate these adversarial attacks. While we have showcased results with multiple deep face networks in this research, we have used VGG to report the detection and mitigation results for DeepFool and universal adversarial perturbations since it is the only network for which the authors have provided pre-computed models.

5.1 Adversarial Attacks on Deep Learning based Face Recognition

In this section, we discuss the proposed adversarial distortions that are able to degrade the performance of deep face recognition algorithms. Let $\mathbf{x}$ be the face image input to a deep learning based face recognition algorithm and l be the output class label (in case of identification, it is an identity label and for verification, it is *same* or *different*). An adversarial attack function $a(\cdot)$, when applied to the input face image, falsely changes the predicted identity label. In other words, if $a(\mathbf{x}) = l'$ where, $l \neq l'$, then a is a successful adversarial attack on the network. While adversarial learning has been used in literature to showcase that the function $a(\cdot)$ can be obtained via optimization

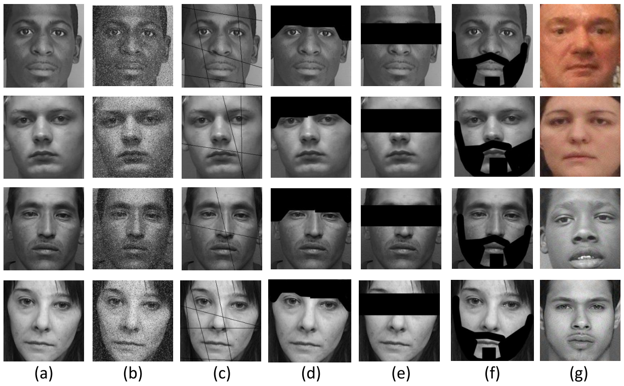


Figure 5-2: Sample images representing the (b) grid based occlusion (Grids), (c) most significant bit based noise (xMSB), (d) forehead and brow occlusion (FHBO), (e) eye region occlusion (ERO), and (f) beard-like occlusion (Beard) distortions when applied to the (a) original images. (g) is the universal perturbed [148] images of PaSC and MEDS databases.

based on network gradients, in this research we explore a different approach. We evaluate the robustness of deep learning based face recognition in the presence of image processing based distortions. Based on the information required in their design, these distortions can be considered at image-level or face-level. We propose two image-level distortions: (a) grid based occlusion, and (b) most significant bit based noise, along with three face-level distortions: (a) forehead and brow occlusion, (b) eye region occlusion, and (c) beard-like occlusion.

5.1.1 Image-level Distortions

Distortions that are not specific to faces and can be applied to an image of any object are categorized as image-level distortions. In this research, we have utilized two such distortions, grid based occlusion and most significant bit change based noise addition. Figure 5-2(b) and 5-2(c) present sample outputs of image-level distortions.

Grid based Occlusion

For the grid based occlusion (termed as Grids) distortion, we select a number of points $P = \{p_1, p_2, ..., p_n\}$ along the upper ($y = 0$) and left ($x = 0$) boundaries of the image according to a parameter ρ_{grids}. The parameter ρ_{grids} determines the number of grids that are used to distort each image with higher values resulting in a denser grid, i.e., more grid lines. For each point $p_i = (x_i, y_i)$, we select a point on the opposite boundary of the image, $p'_i = (x'_i, y'_i)$, with the condition if $y_i = 0$, then $y'_i = H$ and if $x_i = 0$ then $x'_i = W$, where, $W \times H$ is the size of the input image. Once a set of pairs corresponding to points P and P' have been selected for the image, one pixel wide line segments are created to connect each pair, and each pixel lying on these lines is set to 0 grayscale value.

Most Significant Bit based Noise

For the most significant bit based noise (xMSB) distortion, we select three sets of pixels $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ from the image stochastically such that $|\mathcal{X}_i| = \phi_i \times W \times H$, where $W \times H$ is the size of the input image. The parameter ϕ_i denotes the fraction of pixels where the i^{th} most significant bit is flipped. The higher the value of ϕ_i, the more pixels are distorted in the i^{th} most significant bit. For each $\mathcal{P}_j \in X_i, \forall i \in [1, 3]$, we perform the following operation:

$$\mathcal{P}_{kj} = \mathcal{P}_{kj} \oplus 1 \tag{5.1}$$

where, $\mathcal{P}_{kj}$ denotes the k^{th} most significant bit of the j^{th} pixel in the set and $\oplus$ denotes the bitwise XOR operation. It is to be noted that the sets $\mathcal{X}_i$ are not mutually exclusive and may overlap. Therefore, the total number of pixels affected by the noise is at most $|\mathcal{X}_1 + \mathcal{X}_2 + \mathcal{X}_3|$ but may also be lower depending on the stochastic selection.

5.1.2 Face-level Distortions

Face-level distortions specifically require face-specific information, e.g. location of facial landmarks. The three face-level region based occlusion distortions are applied after performing automatic face and facial landmark detection. In this research, we have utilized the open source DLIB library [100] to obtain the facial landmarks. Once facial landmarks are identified, they are used along with their boundaries for masking. To obscure the eye region, a singular occlusion band is drawn on the face image as follows:

$$I\{x, y\} = 0, \forall x \in [0, W], y \in \left[y_e - \frac{d_{eye}}{\psi}, y_e + \frac{d_{eye}}{\psi}\right] \tag{5.2}$$

Here, $y_e = \left(\frac{y_{le}+y_{re}}{2}\right)$, and (x_{le}, y_{le}) and (x_{re}, y_{re}) are the locations of the left eye center and the right

Table 5.2: Characteristics of the databases used for adversarial attack generation and detection.

Database	Subjects	Images
PaSC [12]	293	4,688
MEDS-II [2]	518	858

Table 5.3: Verification performance of existing face recognition algorithms in the presence of different distortions on the PaSC database. All values indicate genuine accept rate (%) at 1% false accept rate.

System	PaSC					
	Original	Grids	xMSB	FHBO	ERO	Beard
COTS	24.1	20.9	14.5	19.0	0.0	24.8
OpenFace	66.7	49.5	43.8	47.9	16.4	48.2
VGG-Face	78.4	50.3	45.0	25.7	10.9	47.7
LightCNN	89.3	80.1	71.5	62.8	26.7	70.7
L-CSSE	89.1	81.9	83.4	55.8	27.3	70.5

eye center, respectively. The inter-eye distance d_{eye} is calculated as: $x_{re} - x_{le}$ and ψ is a parameter that determines the width of the occlusion band. Similar to the eye region occlusion (ERO), the forehead and brow occlusion (FHBO) is created where facial landmarks on forehead and brow regions are used to create a mask. For the beard-like occlusion, outer facial landmarks along with nose and mouth coordinates are utilized to create the mask as combinations of individually occluded regions. Figure 5-2 (d), (e), and (f) illustrate the samples of face-level distortions.

5.1.3 Learning based Adversaries

Along with the proposed image-level and face-level distortions, we also analyze the effect of adversarial samples generated using two existing adversarial models: DeepFool [149] and Universal Adversarial Perturbations [148].

5.2 Adversarial Distortions: Results and Analysis

In this section, we first provide a brief overview of the deep face recognition networks, databases, and respective experimental protocols that are used to conduct the face verification evaluations. We attempt to assess how the deep networks perform in the presence of different kinds of proposed distortions to emphasize the need for addressing such attacks.

Table 5.4: Verification performance of existing face recognition algorithms in the presence of different distortions on the MEDS database. All values indicate genuine accept rate (%) at 1% false accept rate.

System	MEDS					
	Original	Grids	xMSB	FHBO	ERO	Beard
COTS	40.3	24.3	19.1	13.0	0	6.2
OpenFace	39.4	10.1	10.1	14.9	6.5	22.6
VGG-Face	54.3	3.2	1.3	15.2	8.8	24.0
LightCNN	60.1	24.6	29.5	31.9	24.4	38.1
L-CSSE	61.2	43.1	36.9	29.4	39.1	39.8

5.2.1 Databases

We use two publicly available face databases for our experiments, namely, the PaSC database [12] and the MEDS [2] . The PaSC database contains still-to-still and video-to-video matching protocols. We use the frontal subset of the still-to-still protocol which contains 4,688 images pertaining to 293 individuals which are divided into equally sized target and query sets. Each image in the target set is matched to each image in the query set and the resulting 2344×2344 score matrix is used to determine the verification performance.

The MEDS database contains a total of 1,309 faces pertaining to 518 individuals. Similar to the case of PaSC, we utilize the metadata provided with the MEDS release 2 database to obtain a subset of 858 frontal face images from the database. Each of these images is matched to every other image and the resulting 858×858 score matrix is utilized to evaluate the verification performance. For evaluating performance under the effect of distortions, we randomly select 50% of the total images from each database and corrupt them with the proposed distortions separately. These distorted sets of images are utilized to compute the new score matrices for each case.

5.2.2 Existing Networks and Systems

In this research, we utilize the OpenFace [7], VGG-Face [160], LightCNN [218], and L-CSSE [135] networks to gauge the performance of deep face recognition algorithms in the presence of the aforementioned distortions. The OpenFace library is an open source implementation of Facenet [183] and is openly available to all members of the research community for modification and experimental usage. The VGG deep face network is a deep CNN with 11 convolutional blocks where each convolution layer is followed by non-linearities such as ReLU and max pooling. LightCNN is another publicly available deep network architecture for face recognition that is a CNN with maxout activations in each convolutional layer and achieves good results with just five convolutional layers. L-CSSE is a supervised autoencoder formulation that utilizes a class sparsity based supervision penalty in the loss function to improve the classification capabilities of autoencoder based deep networks. In order to assess the relative performance of deep face recognition with a non-deep learning based approach, we compare the performance of these deep learning based algorithms with a COTS matcher. No fine-tuning is performed for any of these algorithms before evaluating their performance on the test databases.

5.2.3 Results and Analysis

Tables 5.3 and 5.4 summarize the effect of image processing based adversarial distortions on OpenFace, VGG-Face, LightCNN, L-CSSE, and COTS. On the PaSC database, as shown in Table 5.3, while OpenFace and COTS perform comparably to each other at about 1% FAR, OpenFace performs better than the COTS algorithm at all further operating points when no distortions are present. However, we observe a sharp drop in OpenFace performance when any distortion is introduced in the data. For instance, with grids attack, at 1% FAR, the GAR of OpenFace drops by 29.3% and of VGG by 28.1%, whereas the performance of COTS only drops by 16% which is about half the drop compared to what OpenFace and VGG experience. We notice a similar scenario in the presence of noise attack where the performance of OpenFace and VGG drops down by about 29% as opposed to loss of 21.2% observed by COTS. In cases of LightCNN and L-CSSE, they both have shown higher performance with original images; however, as shown in Tables 5.3 and 5.4, similar level of drops are observed. It is to be noted that for xMSB and grid attack, L-CSSE is able to achieve relatively better performance because L-CSSE is a supervised version of autoencoder which can handle *noise* better. Overall, deep learning based algorithms experience higher performance drop as opposed to the non-deep learning based COTS. In the case of occlusions, however, deep learning based algorithms suffer less as compared to COTS. It is our assessment that the COTS algorithm fails to perform accurate recognition with the highly limited facial region available in the low-resolution PaSC images in the presence of occlusions. Similar performance

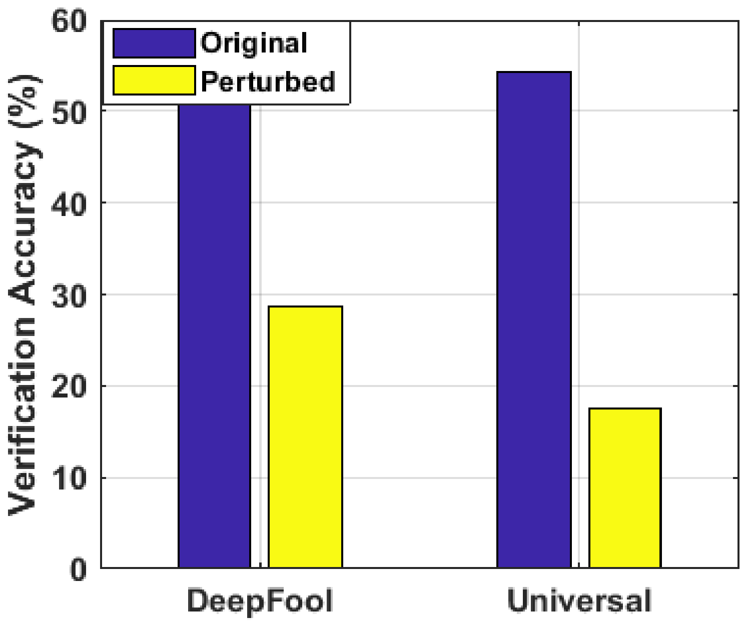


Figure 5-3: Bar graph showing the effect of perturbation on the VGG model. Verification accuracy is reported at 1% GAR.

trends are observed on the MEDS database on which for original images, deep learning based algorithms outperform the COTS matcher with a GAR of 60-89% at 1% FAR respectively as opposed to 24.1% by COTS. The accuracy of deep learning algorithms drops significantly more than the accuracy of COTS.

We next performed a similar analysis with learning based adversaries on the PaSC database. The results of VGGFace model with original and perturbed images are shown in Figure 5-3. It is interesting to observe that the drop in accuracy obtained by simple image processing operations is equivalent to the reduction achieved by learnt adversaries. This clearly shows that deep models are not resilient to even simple perturbations and therefore, it is very important to devise effective strategies for detection and mitigation of attacks.

5.3 Detection and Mitigation of Adversarial Attacks

As we can see in the previous section, adversarial attacks can substantially reduce the performance of usually accurate deep neural network based face recognition methods. Therefore, it is essential to address such singularities in order to make face recognition algorithms more robust and useful

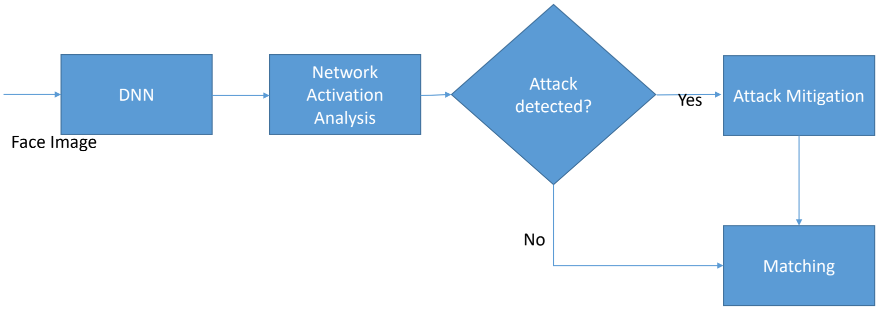


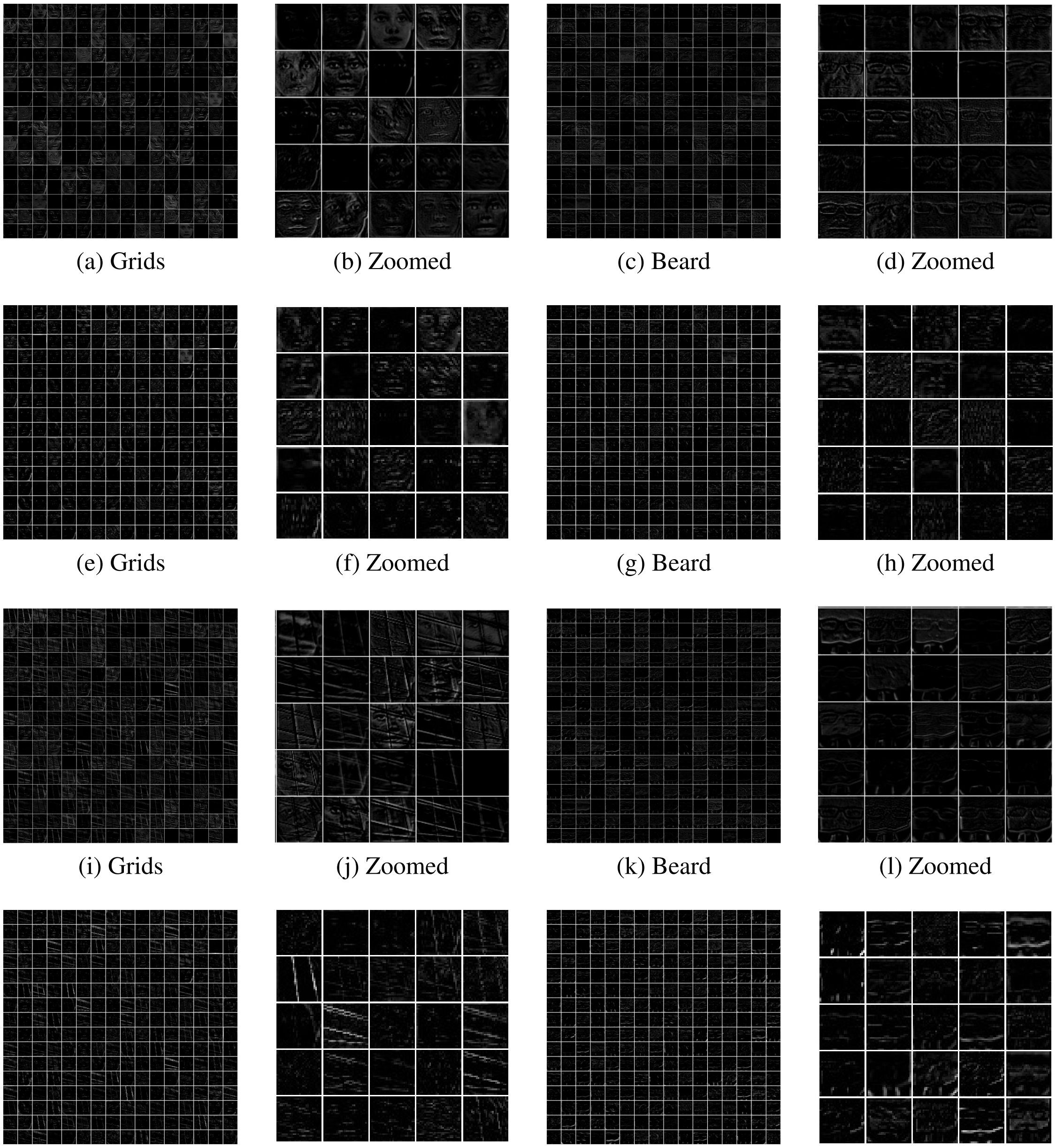
Figure 5-4: Flow chart for the proposed detection and mitigation methodology.

in real world applications. In this section, we propose novel methodologies for detecting and mitigating adversarial attacks. First, we provide a brief overview of a deep network followed by the proposed algorithms and their corresponding results.

Each layer in a deep neural network essentially learns a function or representation of the input data. The final feature computed by a deep network is derived from all of the intermediate representations in the hidden layers. In an ideal scenario, the internal representation at any given layer for an input image should not change drastically with minor changes to the input image. However, that is not the case in practice as proven by the existence of adversarial examples. The final features obtained for a distorted and undistorted image are measurably different from one another since these features map to different classes. Therefore, it is implied that the intermediate representations also vary for such cases. It is our assertion that the internal representations computed at each layer are different for distorted images as compared to undistorted images. Therefore, in order to detect whether an incoming image is perturbed in an adversarial manner, we decide that it is distorted if its layer-wise internal representations deviate substantially from the corresponding mean representations. The overall flow of the detection and mitigation algorithms is summarized in Figure 5-4.

5.3.1 Network Analysis and Detection

In order to develop adversarial attack detection mechanism, we first analyze the filter responses in CNN architecture. Visualizations presented in Figure 5-5 showcase the filter responses for a distorted image at selected intermediate layers that demonstrate the most sensitivity towards noisy data. We can see that many of the filter outputs primarily encode the noise instead of the input



(a) Grids (b) Zoomed (c) Beard (d) Zoomed

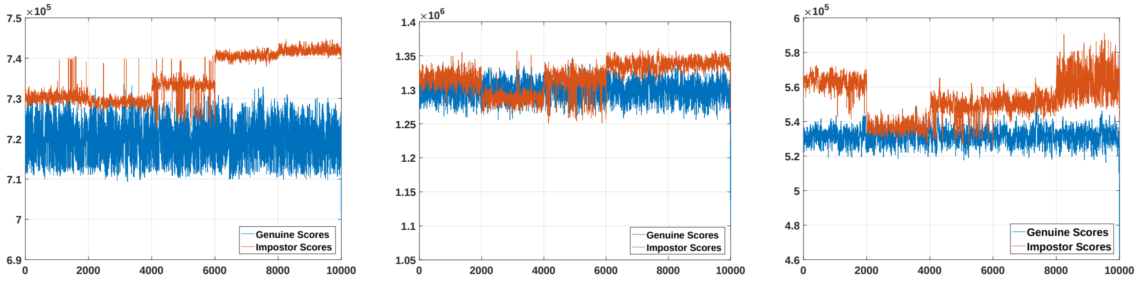
(e) Grids (f) Zoomed (g) Beard (h) Zoomed

(i) Grids (j) Zoomed (k) Beard (l) Zoomed

(m) Grids (n) Zoomed (o) Beard (p) Zoomed

Figure 5-5: Visualizing filter responses for selected layers from the VGG network when the input image is unaltered and affected by the grids and beard distortions. The first two rows present visualizations for conv3_2 and pool3 layers for the original input images respectively. The next two rows present visualizations for the same layers when the input images are distorted using adversarial perturbations. The propagation of the adversarial signal into the intermediate layer representations is the inspiration for our proposed detection and mitigation methodologies.

signal. We observe that the deep network based representation is more sensitive to the input and while that sensitivity results in a more expressive representation that offers higher performance in



(a) Layer 3 (b) Layer 5 (c) Layer 7

Figure 5-6: Visualizing the distribution of genuine (undistorted) and impostor (distorted) sample scores in the Canberra distance space. The scores are obtained by comparing the intermediate layer outputs for the sample with the layer mean obtained using the undistorted training data for a particular layer. In this illustration, we showcase the distributions for scores obtained using layers 3, 5, and 7 of the VGG network for the Multi-PIE database before normalization.

case of undistorted data, it also compromises the robustness towards noise such as the proposed distortions. Since each layer in a deep network learns increasingly more complicated functions of the input data based on the functions learned by the previous layer, any noise in the input data is also encoded in the features thus leading to a higher reduction in the discriminative capacity of the final learned representation. Similar conclusions can also be drawn from the results of other existing adversarial attacks on deep networks, where the addition of a noise pattern leads to spurious classification [60].

To counteract the impact of such attacks and ensure practical applicability of deep face recognition, the networks must either be made more robust towards noise at a layer level during training or it must be ensured that any input is preprocessed to filter out any such distortion prior to computing its deep representation for recognition.

In order to detect distortions we compare the pattern of the intermediate representations for undistorted images with distorted images at each layer. The differences in these patterns are used to train a classifier that can categorize an unseen input as an undistorted/distorted image. In this research, we use the VGG [160] and LightCNN [218] networks to devise and evaluate our detection methodology. From the 50,248 frontal face images in the CMU Multi-PIE database [69], 40,000 are randomly selected and used to compute a set of layer-wise mean representations, μ, as follows:

$$\mu_i = \frac{1}{N_{train}} \Sigma_{j=1}^{N_{train}} \phi_i \left(I_j \right) \qquad (5.3)$$

where, I_j is the j^{th} image in the training set, N_{train} is the total number of training images, μ_i is the mean representation for the i^{th} layer of the network, and $\phi_i(I_j)$ denotes the representation obtained at the i^{th} layer of the network when I_j is the input. Once μ is computed, the intermediate representations computed for an arbitrary image I can be compared with the layer-wise means as follows:

$$\Psi_i(I,\mu) = \Sigma_z^{\lambda_i} \frac{|\phi_i(I)_z - \mu_{iz}|}{|\phi_i(I)_z| + |\mu_{iz}|} \tag{5.4}$$

where, $\Psi_i(I,\mu)$ denotes the Canberra distance between $\phi_i(I)$ and μ_i, λ_i denotes the length of the feature representation computed at the i^{th} layer of the network, and μ_{iz} denotes the z^{th} element of μ_i. If the number of intermediate layers in the network is N_{layers}, we obtain N_{layers} distances for each image I. As illustrated in Figure 5-6, the undistorted and distorted samples are well separated in the Canberra distance score space. These distances are normalized using min-max normalization and then used as features to train a SVM [195] with the RBF kernel for two-class classification. The kernel parameters for the SVM are optimized on the training data using a stochastic grid search methodology.

5.3.2 Mitigation: Selective Dropout

An ideal automated solution should not only automatically detect but also mitigate the effect of an adversarial attack so as to maintain as high performance as possible. Therefore, the next step in defending against adversarial attack is mitigation. This can be achieved by discarding or preprocessing (e.g. denoising) the affected regions. In order to accomplish these objectives, we again utilize the characteristics of the output produced in the intermediate layers of the network. We select 10,000 images from the Multi-PIE database that are partitioned into 5 mutually exclusive and exhaustive subsets of 2,000 images each. Each subset is processed using a different distortion. The set of 10,000 distorted images thus obtained contains 2,000 images pertaining to each of the five proposed distortions. We use a smaller separate Multi-PIE subset of 1,680 faces (5 per subject) for training the algorithm on DeepFool and universal perturbations. Using this data, we compute a filter-wise score per layer that estimates the particular filter's sensitivity towards distortion as follows:

$$\epsilon_{ij} = \Sigma_{k=1}^{N_{dis}} \|\phi_{ij}(I_k) - \phi_{ij}(I_k^{'})\| \tag{5.5}$$

where, N_{dis} is the number of distorted images in the training set, ϵ_{ij} denotes the score and $\phi_{ij}(\cdot)$ denotes the response of the j^{th} filter in the i^{th} layer, I_k is the k^{th} distorted image in the dataset, and $I_k^{'}$ is the undistorted version of I_k. Once these values are computed, the top η layers are selected based on the aggregated ϵ values for each layer. These are the layers identified to contain the most filters that are adversely affected by the distortions in data. For each of the selected η layers, the top κ fraction of affected filters are disabled by modifying the weights pertaining to 0 before computing the features. We also apply a median filter of size 5×5 for denoising the image before extracting the features. We term this approach as *selective dropout*. It is aimed at increasing the network's robustness towards noisy data by removing the most problematic filters from the pipeline. We determine the values of parameters η and κ via grid search optimization on the training data with verification performance as the criterion.

5.3.3 Experimental Protocol

For training the detection model, we use the remaining 10,000 frontal face images from the CMU Multi-PIE database as undistorted samples. We generate 10,000 distorted samples using all five distortions with 2,000 images per distortion that are also randomly selected from the CMU Multi-PIE database. We use the same training data for universal perturbations with 10,000 distorted and 10,000 undistorted samples. For DeepFool, we use a subset of 1,680 face images from the CMU Multi-PIE database with 5 images from each of the 336 subjects with both distorted and undistorted versions for training the detection algorithm. Since the VGGFace network has 20 intermediate layers, we obtain a feature vector of size 20 distances for each image. We perform a grid search based parameter optimization using the $20,000 \times 20$ training matrix to optimize and learn the SVM model. For DeepFool, the size of the training data is $3,360 \times 20$. Once the model is learned, any given test image is characterized by the distance vector and processed by the SVM. The score given by the model for the image to belong to the distorted class is used as a distance metric. We observe that the metric thus obtained is able to classify distorted images on

Table 5.5: Performance (accuracy %) of the proposed detection methodology (using LightCNN and VGG as the target networks) compared to two existing detection algorithms. Grids = grid based occlusion, xMSB = most significant bit based noise, FHBO = forehead and brow occlusion, ERO = eye region occlusion, and Beard = beard like occlusion.

Distortion	MEDS				PaSC			
	LightCNN	VGG	[120]	[51]	LightCNN	VGG	[120]	[51]
Beard	**92.2**	86.8	81.2	80.9	89.5	**99.8**	83.4	85.1
ERO	**91.9**	86.0	80.4	80.0	90.6	**99.7**	84.9	84.6
FHBO	**92.9**	84.4	79.8	79.6	81.7	**99.8**	78.3	77.8
Grids	68.4	**84.4**	62.1	62.4	89.7	**99.9**	85.1	85.7
xMSB	**92.9**	85.4	80.2	80.9	93.2	**99.8**	88.2	87.9

unseen databases. The mitigation algorithm is evaluated with both LightCNN and VGG networks on both the PaSC and MEDS databases with the same experimental protocol as used in obtaining the verification results in Section 5.2.

5.3.4 Results and Analysis

First, we present the results of the proposed algorithm in detecting whether an image contains adversarial distortions or not using the VGG and LightCNN networks. Table 5.5 present the results of adversarial attack detection. We choose these two as the model definition and weights are publicly available. Each distortion based subset comprises of a 50% split of distorted and undistorted faces. These are the same sets that have been used for evaluating the performance of the three face recognition systems. As mentioned previously, the model is trained on a separate database which does not have any overlap with the test set.

The proposed detection algorithm performs almost perfectly for the PaSC database with the VGG network and maintains accuracies of 80-90% with the LightCNN network. The lowest performance is observed on the MEDS database (classification accuracy of 68.4% with the LightCNN network). The lower accuracies with the LightCNN can be attributed to the smaller network depth which results in smaller size features to be utilized by the detection algorithm. It is to be noted that the proposed algorithm maintains high true positive rates even at very low false positive rates across all distortions on both databases which is desirable when the cost of accepting a distorted image is much higher than a false reject for the system. We also observe that the quality based algorithms struggle with high resolution distorted images and low resolution undistorted images, classifying them as undistorted and distorted respectively. Besides exceptionally poor quality images that are naturally quite distorted, we observe that high or low illumination results in false rejects by the algorithm, i.e., falsely detected as distorted. This shows the scope of further improvement and refinement in the detection methodology. This is also another reason for lower performance with the MEDS database which has more extreme illumination cases as compared to PaSC. We observe both general no-reference image quality measures and face-specific quality measures to also be insufficient for attack detection. We also test using the Viola Jones face detector [208] and find that, on average, approximately 60% of the distorted faces pass face detection. Therefore, the distorted face images cannot be differentiated from undistorted faces on the basis of failing face detection. We attempt to reduce the feature dimensionality to deduce the most important features using sequential feature selection based on classification loss by a SVM model learned on a given subset of features. For the VGG based model, using just the top 6 features for detection, we obtain an average accuracy of 81.7% on MEDS and 96.9% on PaSC database across all distortions. If we use only one most discriminative feature to perform detection, we obtain 79.3% accuracy on MEDS and 95.8% on PaSC on average across all distortions. This signifies that comparing the representations computed by the network in its intermediate layers indeed produces a good indicator of the existence of distortions in a given image.

In addition to the proposed adversarial attacks, we have also evaluated the efficacy of the proposed detection methodology on two existing attacks that utilize network architecture information for adversarial perturbation generation, i.e., DeepFool [149] and Universal adversarial perturbations [148]. We have also compared the performance of the proposed detection algorithm with two recent adversarial detection techniques based on adaptive noise reduction [120] and Bayesian uncertainty [51]. Same training data and protocol was used to train and test all three detection approaches. The results of detection are presented in Table 5.5 and Figure 5-7. We observe that the proposed methodology is at least 11% better at detecting image processing based adversarial attacks as compared to the existing algorithms for all cases except for detecting DeepFool perturbed images from the MEDS database where it still outperforms the other approaches by more than 3%. We believe that this is due to the fact that MEDS has overall higher image quality as compared to PaSC and even the impact of these near imperceptible perturbations (DeepFool and Universal) on verification performance is minimal for the database. Therefore, it is harder to distinguish original

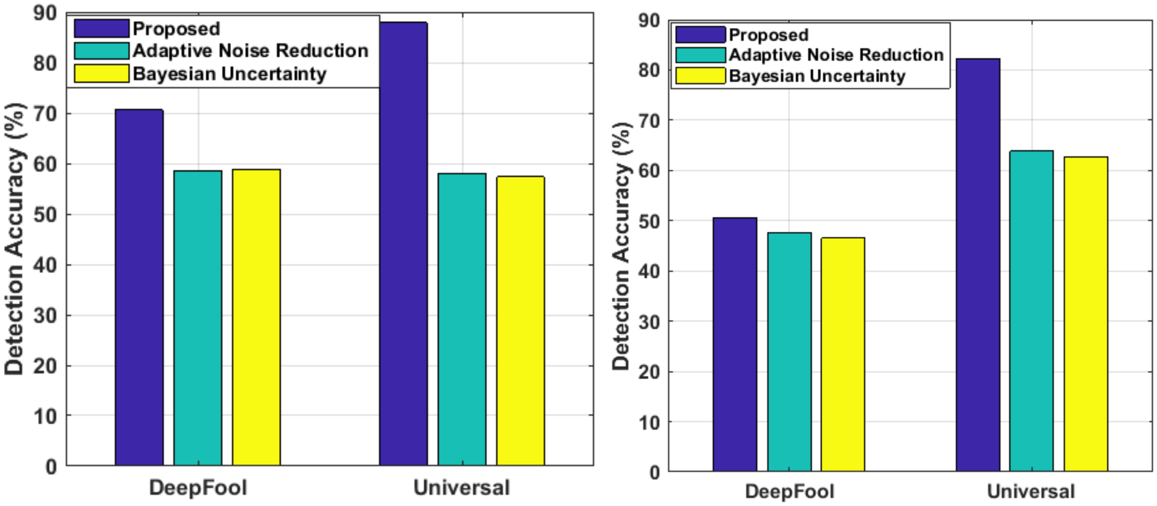


Figure 5-7: Summarizing the results of the proposed and existing detection algorithms on the PaSC and MEDS databases.

Table 5.6: Mitigation Results (GAR (%) at 1% FAR) on the MEDS and PaSC databases.

Algorithm	Database	Original	Distorted	Corrected
LCNN	PaSC	60.5	25.9	**36.2**
	MEDS	89.3	41.6	**61.3**
VGGFace	PaSC	54.3	14.6	**24.8**
	MEDS	78.4	30.5	**40.6**

images from perturbed images for these distortions for all the tested detection algorithms.

Table 5.6 present the results for the mitigation algorithm. Mitigation is a two-step process to enable better performance and computational efficiency. Figure 5-3 shows the effect of deepfool and universal adversary on the verification performance using VGG model. First, using the proposed detection algorithm we perform selective mitigation of only those images that are considered adversarial by the learned model. Face verification results after applying the proposed mitigation algorithm on the MEDS and PaSC databases are presented in Table 5.6. We can observe that the mitigation model is able to improve the verification performance on both the databases with either network and bring it closer to the original. Thus, we see that even discarding a certain fraction of the intermediate network output, that is the most affected by adversarial distortions, results in better recognition than incorporating them into the obtained feature vector.

5.4 Summary

To summarize, our work has three main contributions: (i) a framework to evaluate robustness of deep learning based face recognition engines, (ii) a scheme to detect adversarial attacks on the system; and (iii) methods to mitigate adversarial attacks when detected. Playing the role of an expert level adversary, we propose five classes of image distortions in the evaluation experiment. Using an open source implementation of Facenet, i.e., OpenFace, and the recently proposed VGG-Face, LightCNN, and L-CSSE networks, we conduct a series of experiments on the publicly available PaSC and MEDS databases. We observe a substantial loss in the performance of the deep learning based systems when compared with a non-deep learning based COTS matcher for the same evaluation data. In order to detect the attacks, we propose a network activation analysis based method in the hidden layers of the network. When an attack is reported by this stage, we invoke the described mitigation algorithm to show that we can recover from the attacks in many situations.

Chapter 6

Conclusions and Future Work

This dissertation focuses on different aspects of feature representations for efficient and robust face recognition and tries to bridge some of the gaps that exist in reaching near perfect face recognition in a truly unconstrained environment. There are multiple avenues to improve the performance of face recognition algorithms including fusion of multiple features, learning feature representations in a data-driven manner, using rich data sources such as a video or RGB-D images, and addressing adversarial attacks that might compromise the algorithm's integrity. This dissertation presents algorithms based on these opportunities demonstrating the merits of attacking the problem of accurate face recognition at the input and algorithm levels:

- We propose a novel RGB-D based face recognition algorithm that uses both texture and attribute features to improve performance using data from consumer level sensors.
- We propose a novel methodology for feature level fusion using group sparse representation. We also propose a kernelization based extension to it and successfully apply it in the domains of RGB-D, cross-resolution, and multi-biometric recognition.
- We propose a mechanism to quantize the feature-richness along with a deep joint representation framework to perform accurate and efficient face recognition in videos.
- We explore the susceptibility of deep networks towards adversarial attacks and propose algorithms to detect and mitigate such attacks.

There still remains a lot of scope for future work and improvement. Improving feature extraction at the input and algorithm level is discussed throughout the dissertation and further improvements in these areas is definitely a potential research direction.

- RGB-D video data can be utilized to further enhance the data sources and newer consumer level RGB-D devices can be deployed to obtain higher fidelity depth data. This data can also be used in conjunction with 3D model generation algorithms for improved robustness towards pose variations.

- Future research in group sparsity based approaches can be focused towards applying the group sparsity principle to various other junctures of fusing feature-like information. For example, layer output aggregation and pooling in deep networks.

- Future research in video face recognition can be directed towards combining the proposed architecture with a CNN to extract complimentary feature representations. Another direction can be to explore the use of feature-richness in filtering large training databases designed for deep network training. Filtering the data on the basis of feature-richness might facilitate guided generation of useful samples while reducing the overall number of required data points.

- There is a requirement for extensive future work in making deep networks robust towards adversarial examples. First is to build more general solutions for detecting and mitigating adversarial attacks that can work with all kinds of networks [5, 93]. Additionally, the reason for the existence of adversarial perturbations should be explored in further detail. This may shed more light in solving the fundamental flaws of data-driven feature representations and make them more generalizable even when trained on a limited set of data. The idea of selective dropout can be explored further to improve the robustness and performance of deep learning architectures [165].

- An “augmentation” phase can be utilized such that additional sources of information can be incorporated with face data to improve the decision accuracy. Figure 6-1 provides an overview of where the augmentation phase can be added in the face recognition pipeline. Additional information such as social context (the likely possibility that an individual will

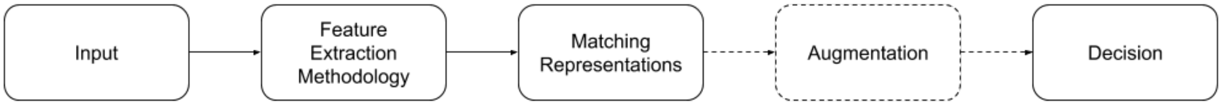


Figure 6-1: Illustrating a potential target for future work to improve face recognition further. The augmentation phase can fit in between the matching and decision phases and supplement face information with other cues that can aid overall recognition.

be present in a photograph with another subject depending on the overlap of their social circles) can be added to the decision process in addition to the comparison of features to further improve the reliability and performance of automated systems. The global context of an image from which the face region of interest has been extracted can also act as an additional signal that can be used to validate the decision of an automated algorithm. Continuous feedback can also be incorporated using the augmentation phase where the recent performance of the algorithm can be utilized as a factor in deciding upon the reliability of future decisions. Spatial and temporal context in the form of location and timestamp information can also be critical in making the right decision in the absence of specialized sensors or multiple data points (such as in surveillance applications). As social networks and mobile devices evolve further, more accurate and rich information about the image/video capture and individuals may become available to the face recognition system which can be included as part of the augmentation phase.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] http://english.cas.cn/Ne/CASE/200808/t20080815_18764.shtml. 89

[2] Multiple Encounters Dataset (MEDS), http://www.nist.gov/itl/iad/ig/sd32.cfm, National Institute of Standards and Technology, 2011. 124, 125

[3] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, R. Nevatia, and G. Medioni. Face recognition using deep multi-pose representations. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–9, 2016. 7, 14

[4] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE International Joint Conference on Biometrics*, pages 659–665, 2017. 119

[5] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha. Are imageagnostic universal adversarial perturbations for face recognition difficult to detect. *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2018. 138

[6] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. 50

[7] B. Amos, B. Ludwiczuk, J. Harkes, P. Pillai, K. Elgazzar, and M. Satyanarayanan. OpenFace: Face Recognition with Deep Neural Networks. http://github.com/cmusatyalab/openface. Accessed: 2016-01-11. 125

[8] Y. Bai, L. Guo, L. Jin, and Q. Huang. A novel feature extraction method using pyramid

histogram of orientation gradients for smile recognition. In *International Conference on Image Processing*, pages 3305–3308, 2009. 36

[9] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008. 50, 68, 72

[10] P. J. Besl and N. D. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 23

[11] J. R. Beveridge, Z. Hao, B. A. Draper, P. J. Flynn, F. Zhenhua, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, M. Kan, R. Wang, S. Shan, X. Chen, H. Li, G. Hua, V. Struc, J. Krizaj, C. Ding, D. Tao, and P. J. Phillips. Report on the FG 2015 Video Person Recognition Evaluation. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8, 2015. 92, 109, 111

[12] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE Conference on Biometrics: Theory, Applications and Systems*, pages 1–8, 2013. XVI, 90, 91, 92, 94, 107, 109, 111, 124, 125

[13] J. R. Beveridge, H. Zhang, P. J. Flynn, Y. Lee, V. E. Liong, J. Lu, M. A. Angeloni, T. F. Pereira, H. Li, G. Hua, V. Struc, J. Krizaj, and P. J. Phillips. The IJCB 2014 PaSC video face and person recognition competition. In *IEEE International Joint Conference on Biometrics*, pages 1–8, 2014. 109, 111

[14] A. N. Bhagoji, D. Cullina, and P. Mittal. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. *arXiv preprint arXiv:1704.02654*, 2017. 121

[15] S. Bharadwaj, H. S. Bhatt, R. Singh, M. Vatsa, and A. Noore. QFuse: Online Learning Framework for Adaptive Biometric System. *Pattern Recognition*, 48(11):3428 – 3439, 2015. 52, 68, 69, 72, 73, 86

[16] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh. Domain specific learning for newborn face recognition. *IEEE Transactions on Information Forensics and Security*, 11(7):1630–1641, 2016. 118

[17] S. Bharadwaj, M. Vatsa, and R. Singh. Biometric quality: a review of fingerprint, iris, and face. *EURASIP Journal on Image and Video Processing*, 2014(1):1–28, 2014. 105, 112

[18] Simone Bianco. Large age-gap face verification by feature injection in deep networks. *Pattern Recognition Letters*, 90:36–42, 2017. XIII, 2

[19] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of SIFT features for face authentication. In *Conference on Computer Vision and Pattern Recognition Workshop*, pages 35–35, 2006. 50, 51

[20] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches to three-dimensional face recognition. In *International Conference on Pattern Recognition*, volume 1, pages 358–361, 2004. 21, 36, 38

[21] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2707–2714, 2010. 8

[22] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 121

[23] Z. Chai, Z. Sun, H. Mendez-Vazquez, R. He, and T. Tan. Gabor ordinal measures for face recognition. *IEEE Transactions on Information Forensics and Security*, 9(1):14–26, 2014. 51

[24] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1160–1165, 2003. 50

[25] J. C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1–9, 2016. 14

[26] J.-C. Chen, R., Ranjan, S. Sankaranarayanan, A. Kumar, C.-H. Chen, V. M. Patel, C. D. Castillo, and R. Chellappa. Unconstrained still/video-based face verification with deep convolutional neural networks. *International Journal of Computer Vision*, 126(2-4):272–291, 2018. 90

[27] M. Chen, A. O'Sullivan, N. Singla, E. J. Sirevaag, S. D. Kristjansson, P.-H. Lai, A. D. Kaplan, and J. W. Rohrbaugh. Laser doppler vibrometry measures of physiological function: Evaluation of biometric capabilities. *IEEE Transactions on Information Forensics and Security*, 5(3):449–460, 2010. 51

[28] Y. Chen, N. M. Nasrabadi, and T. D. Tran. Hyperspectral image classification via kernel sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):217–231, 2013. 57

[29] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *European Conference on Computer Vision*, pages 766–779, 2012. 90

[30] P. Chhokra, A. Chowdhury, G. Goswami, M. Vatsa, and R. Singh. Unconstrained kinect video face database. *Information Fusion*, 44:113 – 125, 2018. XIII, 2

[31] Y.J. Chin, T.S. Ong, A.B.J. Teoh, and K.O.M. Goh. Integrated biometrics template protection technique based on fingerprint and palmprint feature-level fusion. *Information Fusion*, 18(0):161 – 174, 2014. 51

[32] E. Corvee and F. Bremond. Body parts detection for people tracking using trees of histogram of oriented gradient descriptors. In *Advanced Video and Signal-Based Surveillance*, pages 469–475, 2010. 29

[33] D. Cox and N. Pinto. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 8–15, 2011. 7, 9

[34] S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak. A Protocol for Multibiometric Data Acquisition, Storage and Dissemination. Technical report, WVU, 2007. 52, 68, 85

[35] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005. 24, 25, 29, 51, 79

[36] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. *arXiv preprint arXiv:1705.02900*, 2017. 121

[37] R. L. de Carvalho and P. F. F. Rosa. Identification system for smart homes using footstep sounds. In *IEEE International Symposium on Industrial Electronics*, pages 1639–1644, 2010. 50

[38] W. Deng, J. Hu, and J. Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1864–1870, 2012. 9

[39] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Reviews of Neuroscience*, 18(1):193–222, 1995. 27

[40] C. Ding, J. Choi, D. Tao, and L. S. Davis. Multi-directional multi-level dual-cross patterns for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):518–531, 2016. 13

[41] C. Ding and D. Tao. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, 17(11):2049–2058, 2015. 12

[42] C. Ding and D. Tao. Trunk-Branch Ensemble Convolutional Neural Networks for Video-based Face Recognition. *CoRR*, abs/1607.05427, 2016. 91, 92, 116, 117

[43] C. Ding and D. Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 66:144–152, 2017. 92

[44] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):1002–1014, 2018. 92

[45] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24(3):980–993, 2015. 12

[46] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama. 3D face recognition under expressions, occlusions, and pose variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2270–2283, 2013. 21

[47] E. Elhamifar and R. Vidal. Robust classification using structured sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1873–1879, 2011. 55, 56

[48] S. Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637):68–71, 1997. 27

[49] N. Engelhard, F. Endres, J. Hess, J. Sturm, and W. Burgard. Real-time 3D visual slam with a hand-held RGB-D camera. In *RGB-D Workshop, European Robotics Forum*, 2011. 22

[50] Z. Fan, D. Zhang, X. Wang, Q. Zhu, and Y. Wang. Virtual dictionary based kernel sparse representation for face recognition. *Pattern Recognition*, 76:1 – 13, 2018. 17

[51] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017. 121, 133, 134

[52] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 29

[53] G. E. Hinton, and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 102

[54] G. E. Hinton and R. Salakhutdinov. A better way to pretrain deep boltzmann machines. In *Advances in Neural Information Processing Systems*, volume 25, pages 2447–2455. 2012. 100, 105

[55] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, and D. Yue. Learning robust and discriminative low-rank representations for face recognition with occlusion. *Pattern Recognition*, 66:129 – 143, 2017. 16

[56] S. Gao, I. W. H. Tsang, and L. T. Chia. Sparse representation with kernels. *IEEE Transactions on Image Processing*, 22(2):423–434, 2013. 53

[57] Y. Gao, J. Ma, and A. L. Yuille. Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples. *IEEE Transactions on Image Processing*, 26(5):2545–2560, 2017. 16

[58] Soumyadeep Ghosh, Tejas I Dhamecha, Rohit Keshari, Richa Singh, and Mayank Vatsa. Feature and keypoint selection for visible to near-infrared face matching. In *IEEE International Conference on Biometrics Theory, Applications and Systems, 2015*, pages 1–7, 2015. 1

[59] Z. Gong, W. Wang, and W.-S. Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017. 121

[60] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 119, 121, 130

[61] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh. On RGB-D face recognition using kinect. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2013. 18, 30

[62] G. Goswami, R. Bhardwaj, R. Singh, and M. Vatsa. MDLFace: Memorability augmented deep learning for video face recognition. In *International Joint Conference on Biometrics*, pages 1–7, 2014. 92, 94, 111, 114

[63] G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh. Group sparse representation based classification for multi-feature multimodal biometrics. *Information Fusion*, 32:3–12, 2016. 19

[64] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. *AAAI*, 2018. 20

[65] G. Goswami, R. Singh, M. Vatsa, and A. Majumdar. Kernel group sparse representation based classifier for multimodal biometrics. In *International Joint Conference on Neural Networks*, pages 2894–2901, 2017. 19

[66] G. Goswami, M. Vatsa, and R. Singh. RGB-D face recognition with texture and attribute features. *IEEE Transactions on Information Forensics and Security*, 9(10):1629–1640, 2014. XX, 51, 79, 84, 85

[67] G. Goswami, M. Vatsa, and R. Singh. Face verification via learned representation on feature-rich video frames. *IEEE Transactions on Information Forensics and Security*, 12:1686–1698, 2017. 7, 18, 20

[68] M. Grgic, K. Delac, and S. Grgic. Scface — surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, 2011. XIII, 2, 80

[69] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. XIII, 2, 130

[70] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017. 121

[71] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji. Discriminant sparse neighborhood preserving embedding for face recognition. *Pattern Recognition*, 45(8):2884 – 2893, 2012. 9

[72] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3499–3506, 2013. 90

[73] H. Mendez-Vazquez, Y. Martinez-Diaz, and Z. Chai. Volume structured ordinal features with background similarity measure for video face recognition. In *International Conference on Biometrics*, 2013. 91

[74] H. S. Bhatt, R. Singh, and M. Vatsa. On recognizing faces in videos using clustering based re-ranking and fusion. *IEEE Transactions on Information Forensics and Security*, 9(7):1056 – 1068, 2014. 90, 91

[75] M. Hayat, S. H. Khan, and M. Bennamoun. Empowering simple binary classifiers for image set based face recognition. *International Journal of Computer Vision*, pages 1–20, 2017. 15

[76] R. He, T. Tan, L. Davis, and Z. Sun. Learning structured ordinal measures for video based face recognition. *Pattern Recognition*, pages 4–14, 2017. 16

[77] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In *International Symposium on Experimental Robotics*, volume 20, pages 22–25, 2010. 22

[78] R. I. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet. An RGB-D database using microsoft's kinect for windows for face detection. In *International Conference on Signal Image Technology and Internet Based Systems*, pages 42–46, 2012. 22, 34, 35

[79] G. E. Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11:428–434, 2007. 105

[80] T. K. Ho. Random decision forests. In *International Conference on Document Analysis and Recognition*, pages 278–282, 1995. 30

[81] D. Holz, S. Holzer, R. Rusu, and S. Behnke. Real-time plane segmentation using RGB-D cameras. *RoboCup 2011*, pages 306–317, 2012. 22

[82] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *Asian Conference on Computer Vision*, pages 1–16, 2014. 91

[83] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 7, 12, 14

[84] K.-K. Huang, D.-Q. Dai, C.-X. Ren, Y.-F. Yu, and Z.-R. Lai. Fusing landmark-based features at kernel level for face recognition. *Pattern Recognition*, 63:406 – 415, 2017. 16

[85] Z. Huang, W. Li, J. Wang, and T. Zhang. Face recognition based on pixel-level and feature-level fusion of the top-level's wavelet sub-bands. *Information Fusion*, 22(0):95 – 104, 2015. 51

[86] Z. Huang, Y. Liu, C. Li, M. Yang, and L. Chen. A robust face and ear based multimodal biometric system using sparse representation. *Pattern Recognition*, 46(8):2156–2168, 2013. 52

[87] Z. Huang, Y. Liu, X. Li, and J. Li. An adaptive bimodal recognition framework using sparse coding for face and ear. *Pattern Recognition Letters*, 53(0):69 – 76, 2015. 51

[88] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 92

[89] T. Huynh, R. Min, and J. L. Dugelay. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In *Asian Conference on Computer Vision*, 2012. 22, 34, 35

[90] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 27

[91] J. Hu, J. Lu, and Y. Tan. Discriminative deep metric learning for face verification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875 – 1882, 2014. 90, 91

[92] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas. Face recognition from video: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(05), 2012. 89

[93] A. Jain, R. Singh, and M. Vatsa. On detecting gans and retouching based synthetic alterations. In *IEEE Conference on Biometrics: Theory, Applications, and Systems*, 2018. 138

[94] A. K. Jain, P. Flynn, and A. Ross. *Handbook of biometrics*. Springer, 2007. 49, 69, 72

[95] A. K. Jain, L. Hong, and R. Bolle. On-line fingerprint verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):302–314, 1997. 50

[96] A. K. Jain and S. Z. Li. *Handbook of face recognition*. Springer, 2005. 21

[97] R. R. Jillela and A. Ross. Adaptive frame selection for improved face recognition in low-resolution videos. In *International Joint Conference on Neural Networks*, pages 1439–1445, 2009. 95

[98] K. Lee, J. Ho, M. Yang and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 99(3):303–331, 2005. 90

[99] N. M. Khan, N. Xiaoming, A. Quddus, E. Rosales, and L. Guan. On video based face recognition through adaptive sparse dictionary. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015. 91

[100] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 123

[101] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015. 14

[102] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2973–2980, 2012. 31

[103] K. Krishneswari and S. Arumugam. Multimodal biometrics using feature fusion. *Journal of Computer Science*, 8(3):431–435, 2012. 50

[104] A. Kumar and T.-S. Chan. Iris recognition using quaternionic sparse orientation code (qsoc). In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 59–64, 2012. 51, 52, 61

[105] A. Kumar, T.-S. Chan, and C.-W. Tan. Human identification from at-a-distance face images using sparse representation of local iris features. In *IAPR International Conference on Biometrics*, pages 303–309, 2012. 52

[106] A. Kumar and A. Passi. Comparison and combination of iris matchers for reliable personal authentication. *Pattern Recognition*, 43(3):1016 – 1026, 2010. 1

[107] A. Kumar and S. Shekhar. Personal identification using multibiometrics rank-level fusion. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, 41(5):743–752, 2011. 33

[108] A. Kumar, D. C. M. Wong, H. C. Shen, and A. K. Jain. Personal verification using palmprint and hand geometry biometric. In *Audio-and Video-Based Biometric Person Authentication*, pages 668–678, 2003. 50

[109] A. Kumar and C. Wu. Automated human identification using ear imaging. *Pattern Recognition*, 45(3):956–968, 2012. 33

[110] A. Kumar and Y. Zhou. Human identification using finger images. *IEEE Transactions on image processing*, 21(4):2228–2244, 2012. 33

[111] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *International Conference on Computer Vision*, 2009. 31

[112] D. Kun, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Computer Vision and Pattern Recognition*, pages 3474–3481, 2012. 31

[113] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 121

[114] L. Wolf, T. Hassner and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534, 2011. XVI, 7, 9, 12, 90, 92, 93, 94, 98, 107

[115] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Workshop on the Applications of Computer Vision*, pages 186–192, 2013. 10, 22

[116] H. Li and G. Hua. Hierarchical-PEP model for real-world face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 91

[117] H. Li and G. Hua. Probabilistic elastic part model: A pose-invariant representation for real-world face verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. 7, 15

[118] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-PEP for video face recognition. In *Asian Conference on Computer Vision*, pages 1–16, 2014. 91

[119] J. Li and C.-Y. Lu. A new decision rule for sparse representation based classification for face recognition. *Neurocomputing*, 116:265–271, 2013. 54, 55, 62

[120] B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang. Detecting adversarial examples in deep networks with adaptive noise reduction. *CoRR*, abs/1705.08378, 2017. 121, 133, 134

[121] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1193–1205, 2013. 10

[122] J. Liu, Y. Deng, T. Bai, and C. Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *CoRR*, abs/1506.07310, 2015. 13, 119

[123] L. Liu, B. Liu, H. Huang, and A. C. Bovik. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, 29(8):856–863, 2014. 94, 111, 114

[124] W. Liu, Z. Li, and X. Tang. Spatio-temporal embedding for statistical face recognition from video. In *European Conference on Computer Vision*, pages 374–388. 2006. 95

[125] X. Liu, L. Lu, Z. Shen, and K. Lu. A novel face recognition algorithm via weighted kernel sparse representation. *Future Generation Computer Systems*, 80:653 – 663, 2018. 17

[126] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999. 36

[127] C. Lu and X. Tang. Surpassing human-level face verification performance on lfw with gaussianface. In *AAAI*, pages 3811–3819, 2015. 1

[128] J. Lu, T. Issaranon, and D. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. *arXiv preprint arXiv:1704.00103*, 2017. 121

[129] J. Lu, V. E. Liong, G. Wang, and P. Moulin. Joint feature learning for face recognition. *IEEE Transactions on Information Forensics and Security*, 10(7):1371–1383, 2015. 13

[130] J. Lu, V. E. Liong, X. Zhou, and J. Zhou. Learning compact binary face descriptor for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2041–2056, 2015. 13

[131] J. Lu, G. Wang, and P. Moulin. Localized multifeature metric learning for image-set-based face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):529–540, 2016. 13

[132] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):331–345, 2014. 51

[133] Z. Lu, X. Jiang, and A. C. Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Processing Letters*, 2018. 17

[134] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2586–2593, June 2012. 10

[135] A. Majumdar, R. Singh, and M. Vatsa. Face recognition via class sparsity based supervised encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1273–1280, 2017. 125

[136] A. Majumdar and R. K. Ward. Improved group sparse classifier. *Pattern Recognition Letters*, 31(13):1959–1964, 2010. 57

[137] A. Majumdar and R. K. Ward. Robust classifiers for data reduced via random projections. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(5):1359–1371, 2010. 55, 56, 61

[138] A. Majumdar and R. K. Ward. Face recognition from video: An MMV recovery approach. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2221–2224, 2012. 58

[139] A. Majumdar, R. K. Ward, and T. Aboulnasr. Generalized non-linear sparse classifier. In *European Signal Processing Conference*, pages 1–5, 2013. 57

[140] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *IEEE Transactions on Information Forensics and Security*, 12(7):1713–1723, 2017. 119

[141] A. Martinez. The AR face database. *CVC Technical Report*, 24, 1998. XIII, 2

[142] I. Masi, F. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, T. Hassner, et al. Learning pose-aware models for pose-invariant face recognition in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 17

[143] I. Masi, A. T. Trážǧn, T. Hassner, J. T. Leksut, and G. Medioni. *European Conference on Computer Vision*, chapter Do We Really Need to Collect Millions of Faces for Effective Face Recognition?, pages 579–596. 2016. 14

[144] D. S. Matovski, M. S. Nixon, S. Mahmoodi, and J. N. Carter. The effect of time on gait recognition performance. *IEEE Transactions on Information Forensics and Security*, 7(2):543–552, 2012. 50

[145] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017. 121

[146] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 94, 111, 114

[147] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind quality analyzer. *Signal Processing Letters*, 20(3):209–212, March 2013. 94, 111, 114

[148] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. XVII, 121, 122, 124, 134

[149] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *arXiv preprint arXiv:1511.04599*, 2015. 121, 124, 134

[150] S. Nagpal, M. Vatsa, and R. Singh. Sketch recognition: What lies ahead? *Image and Vision Computing*, 55:9–13, 2016. 1

[151] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015. 119

[152] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *Asian Conference on Computer Vision*, pages 709–720, 2010. 7, 8

[153] M. S. Nixon and A. S. Aguado. *Feature extraction & image processing for computer vision*. Academic Press, 2012. 2

[154] I. Odinaka, J. A. O'Sullivan, E. J. Sirevaag, and J. W. Rohrbaugh. Cardiovascular biometrics: Combining mechanical and electrical signals. *IEEE Transactions on Information Forensics and Security*, 10(1):16–27, 2015. 51

[155] I. Odinaka, L. Po-Hsiang, A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, and J. W. Rohrbaugh. Ecg biometric recognition: A comparative analysis. *IEEE Transactions on Information Forensics and Security*, 7(6):1812–1824, 2012. 51

[156] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 68

[157] Y. Ouyang, N. Sang, and R. Huang. Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers. *Neurocomputing*, 149(A):71 – 78, 2015. 62

[158] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12):3371–3408, 2010. 98, 99

[159] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy*, pages 372–387, 2016. 121

[160] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 13, 91, 115, 116, 117, 125, 130

[161] T. Pei, L. Zhang, B. Wang, F. Li, and Z. Zhang. Decision pyramid classifier for face recognition under complex variations using single sample per person. *Pattern Recognition*, 64:305 – 313, 2017. 16

[162] C. Peng, X. Gao, N. Wang, and J. Li. Graphical representation for heterogeneous face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):301–312, 2017. 15

[163] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 763–770, 2010. 104

[164] L. Qiao, S. Chen, and X. Tan. Sparsity preserving projections with applications to face recognition. *Pattern Recognition*, 43(1):331 – 341, 2010. 8

[165] M. Vatsa R. Keshari and R. Singh. Guided dropout. In *AAAI Conference on Artificial Intelligence*, 2019. 138

[166] R. Min, N. Kose, and J. L. Dugelay. KinectFaceDB: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(11):1534–1548, 2014. 79, 84

[167] A. Ramey, V. Gonzalez-Pacheco, and M. A. Salichs. Integration of a low-cost RGB-D sensor in a social robot for gesture recognition. In *Human-Robot Interaction*, pages 229–230, 2011. 22

[168] R. Ranjan, A. Bansal, H. Xu, S. Sankaranarayanan, J.-C. Chen, C. D. Castillo, and R. Chellappa. Crystal loss and quality pooling for unconstrained face verification and recognition. *CoRR*, abs/1804.01159, 2018. 91

[169] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi:10.1109/TPAMI.2017.2781233, 2018. 24

[170] Y. Rao, J. Lu, and J. Zhou. Attention-aware deep reinforcement learning for video face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3931–3940, 2017. 92

[171] R. Rathore, S. Prakash, and P. Gupta. Efficient human recognition system using ear and profile face. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, 2013. 50

[172] A. Rattani, D. R. Kisku, M. Bicego, and M. Tistarelli. Feature level fusion of face and fingerprint biometrics. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, 2007. 50

[173] J. Rauber, W. Brendel, and M. Bethge. Foolbox v0.8.0: A python toolbox to benchmark the robustness of machine learning models. *CoRR*, abs/1707.04131, 2017. 121

[174] A. Ross and A. K. Jain. A prototype hand geometry-based verification system. In *Audio and Video Based Biometric Person Authentication*, pages 166–171, 1999. 50

[175] A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer, 2006. 33, 92, 106

[176] A. A. Ross and R. Govindarajan. Feature level fusion of hand and face biometrics. In *Defense and Security*, pages 196–204. International Society for Optics and Photonics, 2005. 50

[177] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 51

[178] A. Rrnyi. On measures of entropy and information. In *Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1961. 25, 94, 96

[179] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009. 99

[180] C. Samir, A. Srivastava, and M. Daoudi. Three-dimensional face recognition using shapes of facial curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1858–1863, 2006. 21

[181] P.-A. Savalle, E. Richard, and N. Vayatis. Estimation of simultaneously sparse and low rank matrices. In *International Conference on Machine Learning*, 2012. 104

[182] W. J. Scheirer, P. J. Flynn, C. Ding, G. Guo, V. Struc, M. A. Jazaery, K. Grm, S. Dobrisek, D. Tao, Y. Zhu, J. Brogan, S. Banerjee, A. Bharati, and B. RichardWebster. Report on the BTAS 2016 Video Person Recognition Evaluation. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–8, 2016. 116

[183] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, June 2015. 7, 12, 91, 115, 119, 125

[184] M. P. Segundo, S. Sarkar, D. Goldgof, and L. Silva O. Bellon. Continuous 3D face authentication using RGB-D cameras. In *Computer Vision and Pattern Recognition Biometrics Workshop*, 2013. 23

[185] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540, 2016. 119

[186] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa. Joint sparse representation for robust multimodal biometrics recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):113–126, 2014. 52

[187] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *British Machine Vision Conference*, volume 2, page 4, 2013. 7, 10

[188] M. Singh, S. Nagpal, N. Gupta, S. Gupta, S. Ghosh, R. Singh, and M. Vatsa. Cross-spectral cross-resolution video database for face recognition. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–7, 2016. 1

[189] R. Singh, M. Vatsa, and A. Noore. Hierarchical fusion of multi-spectral face images for improved recognition performance. *Information Fusion*, 9(2):200–210, 2008. 49, 50

[190] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 103

[191] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015. 12

[192] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014. 7, 12, 119

[193] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 12, 91

[194] Y. Sun, X. Wang, and X. Tang. Sparsifying neural network connections for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 14

[195] J. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999. 3, 131

[196] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 13, 91, 116

[197] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 121

[198] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481. 2004. 51, 72

[199] Y. Taigman, L. Wolf, T. Hassner, et al. Multiple one-shots for utilizing class label information. In *British Machine Vision Conference*, volume 2, pages 1–12, 2009. 7, 8

[200] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650, June 2010. 8

[201] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. arXiv:1612.04904v1, 2016. 13, 91, 116

[202] A. Tsai, A. Yezzi, and A. S. Willsky. Curve evolution implementation of the mumford-shah functional for image segmentation, denoising, interpolation, and magnification. *IEEE Transactions on Image Processing*, 10(8):1169–1186, 2001. 69

[203] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 50

[204] U. Park, A. K. Jain, and A. Ross. Face recognition in video: Adaptive fusion of multiple matchers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 95

[205] M. Vatsa, R. Singh, and A. Noore. Improving iris recognition performance using segmentation, quality enhancement, match score fusion, and indexing. *IEEE Transactions on Systems, Man, and Cybernetics.*, 38(4):1021–1035, 2008. 69

[206] V. Vijayan, K. W. Bowyer, P. J. Flynn, Di Huang, Liming Chen, M. Hansen, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris. Twins 3D face recognition challenge. In *International Joint Conference on Biometrics*, pages 1–7, 2011. XIV, XIX, 22, 46, 47

[207] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001. 77

[208] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 134

[209] C. Wang and L. Guan. Graph cut video object segmentation using histogram of oriented gradients. In *International Symposium on Circuits and Systems*, pages 2590–2593, 2008. 29

[210] S.-J. Wang, J. Yang, N. Zhang, and C.-G. Zhou. Tensor discriminant color space for face recognition. *IEEE Transactions on Image Processing*, 20(9):2490–2501, 2011. 23

[211] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen. Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015. 91

[212] W. Wang, R. Wang, S. Shan, and X. Chen. Discriminative covariance oriented representation learning for face recognition with image sets. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5599–5608, 2017. 92

[213] G. Wen, Y. Mao, D. Cai, and X. He. Split-net: Improving face recognition in one forwarding operation. *Neurocomputing*, 314:94 – 100, 2018. 17

[214] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. *European Conference on Computer Vision*, chapter A Discriminative Feature Learning Approach for Deep Face Recognition, pages 499–515. 2016. 14

[215] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *European Conference on Computer Vision Real Faces Workshop*, 2008. 7, 8, 36, 79

[216] L. Wolf and N. Levy. The svm-minus similarity score for video face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3523–3530, 2013. 90

[217] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. 8, 23, 36, 38, 51, 53, 54, 55, 69

[218] X. Wu, R. He, Z. Sun, and T. Tan. A lightCNN for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2015. 125, 130

[219] M. Xi, L. Chen, D. Polajnar, and W. Tong. Local binary pattern network: A deep learning approach for face recognition. In *IEEE International Conference on Image Processing*, pages 3224–3228, 2016. 7, 13

[220] Y. Bengio, L. Pascal, P. Dan, and H. Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, volume 19, pages 153–160. 2007. 98

[221] Y. Lee and R. J. Micheals and J. J. Filliben and P. J. Phillips. VASIR: An Open-Source Research Platform for Advanced Iris Recognition Technologies. *Journal of Research of NIST*, 118:218–259, 2013. 69

[222] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701 – 1708, 2014. 7, 12, 90, 91, 93, 115, 119

[223] S. Yadav, M. Singh, M. Vatsa, R. Singh, and A. Majumdar. Low rank group sparse representation based classifier for pose variation. In *IEEE International Conference on Image Processing*, pages 2986–2990, 2016. 104

[224] H. Yan, J. Lu, W. Deng, and X. Zhou. Discriminative multimetric learning for kinship verification. *IEEE Transactions on Information Forensics and Security*, 9(7):1169–1178, 2014. 51

[225] J. Yang and C. Liu. Color image discriminant models and algorithms for face recognition. *IEEE Transactions on Neural Networks*, 19(12):2088–2098, 2008. 23

[226] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):156–171, 2017. 15

[227] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural Aggregation Network for Video Face Recognition. *CoRR*, abs/1603.05474, March 2016. 91, 116, 118

[228] Y.-F. Yao, X.-Y. Jing, and H.-S. Wong. Face and palmprint feature level fusion for single sample biometrics recognition. *Neurocomputing*, 70(7):1582 – 1586, 2007. 50

[229] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 14, 116

[230] J. Yin, Z. Liu, Z. Jin, and W. Yang. Kernel sparse representation based classification. *Neurocomputing*, 77(1):120 – 128, 2012. 57, 58

[231] Y. Ying and P. Li. Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 13:1–26, 2012. 7, 10

[232] Y.-F. Yu, D.-Q. Dai, C.-X. Ren, and K.-K. Huang. Discriminative multi-scale sparse coding for single-sample face recognition with occlusion. *Pattern Recognition*, 66:302 – 312, 2017. 16

[233] X. T. Yuan, X. Liu, and S. Yan. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing*, 21(10):4349–4360, 2012. 55, 56, 61

[234] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3554–3561, 2013. 7, 10, 91

[235] B. Zhang, Y. Gao, S. Zhao, and J. Liu. Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor. *IEEE Transactions on Image Processing*, 19(2):533–544, 2010. 8

[236] B. Zhang, S. Shan, X. Chen, and W. Gao. Histogram of gabor phase patterns (hgpp): A novel object representation approach for face recognition. *IEEE Transactions on Image Processing*, 16(1):57–68, 2007. 7

[237] L. Zhang, J. Chen, Y. Lu, and P. Wang. Face recognition using scale invariant feature transform and support vector machine. In *International Conference for Young Computer Scientists*, pages 1766–1770, 2008. 36

[238] L. Zhang, M. Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *International Conference on Computer Vision*, pages 471–478, 2011. 9

[239] L. Zhang, W.-D. Zhou, P.-C. Chang, J. Liu, Z. Yan, T. Wang, and F.-Z. Li. Kernel sparse representation-based classifier. *IEEE Transactions on Signal Processing*, 60(4):1684–1695, 2012. 57

[240] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2691–2698, 2010. 9

[241] Y. Zhang, D. Zhao, J. Sun, G. Zou, and W. Li. Adaptive convolutional neural network and its application in face recognition. *Neural Processing Letters*, 43(2):389–399, 2016. 14

[242] Z. Zhao and A. Kumar. Towards more accurate iris recognition using deeply learned spatially corresponding features. In *Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy*, pages 22–29, 2017. 51

[243] X. Zhou and B. Bhanu. Feature fusion of side face and gait for video-based human identification. *Pattern Recognition*, 41(3):778 – 795, 2008. 50

[244] X. Zhou, J. Hu, J. Lu, Y. Shang, and Y. Guan. Kinship verification from facial images under uncontrolled conditions. In *ACM International Conference on Multimedia*, pages 953–956, 2011. 51

[245] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 104