Pushing Boundaries of Face Recognition: Adversary, Heterogeneity, and Scale

by

Tejas Indulal Dhamecha

Under the supervision of

Dr. Richa Singh

Dr. Mayank Vatsa

Indraprastha Institute of Information Technology Delhi

July, 2017

©Tejas Indulal Dhamecha, 2017.

Pushing Boundaries of Face Recognition: Adversary, Heterogeneity, and Scale

by

Tejas Indulal Dhamecha

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

to the
Indraprastha Institute of Information Technology Delhi
July, 2017

Certificate

This is to certify that the thesis titled "**Pushing Boundaries of Face Recognition: Adversary, Heterogeneity, and Scale**" being submitted by **Tejas Indulal Dhamecha** 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.

July, 2017
Dr. Richa Singh

July, 2017
Dr. Mayank Vatsa

Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgment

I would like to take this opportunity to show my gratitude to a set of important people in my life. I am not sure if the words are entirely capable to show my gratefulness for the tremendous guidance, mentoring, help, care, companionship, and love I have received from these people.

Foremost, I would thank my advisors and mentors Dr. Richa Singh and Dr. Mayank Vatsa. They have been instrumental in nurturing my research appetite. They have played an important role in improving my communication skills too. In the moments of disappointments, they have boosted my confidence and have never failed to keep my optimism alive. I am fortunate to have them as my advisors. On countless incidences you have gone out of your way to help your students. Surely, the Chhoti Diwali that you decided to spend with students, did make me feel at home even though I was more than thousand kilometers away from my biological family.

I am fortunate to have worked under Dr. Afzel Noore during my visit to the West Virginia University. I have benefited from his knowledge and wisdom. He is the kind-hearted human being that I would like to become one day. I would like to thank Dr. Keith Morris from the Department of Forensics and Investigative Science at WVU for introducing me to the interesting areas of forensics. I would like to thank Dr. Ajay Kumar for hosting me as a visiting research scholar at the Hong Kong Polytechnic University. Further, he has provided me opportunities to serve as volunteer and reviewer in various conferences and journals.

Two of my closest friends, Anush and Paridhi, have played an irreplaceably crucial role in my survival during the doctoral program. I have always found them by my side when I have needed them; I have also found them by my side when I did not actually realize how critically I needed them. I will miss the frustration-releasing PizzaHut dinners with them.

Countless number of brainstorming sessions with Samarth and Himanshu have been very helpful. I owe my PPT skills to Samarth. He has helped me calm down, in general. Shahenshah, as we fondly call Himanshu, has set an example of hard-work, meticulousness, and an awe-inspiring list of publications.

Anupama, Denzil, Kuldeep, Amit ji, Sangeeta, Samy, Nilesh ji, Niharika, Siddhartha, Aayushi, Monika have been very kind and helpful; without their support my stay at IIITD might not have been this smoother. It is during my PhD that I have met some of the sweetest, kindest, and some-

times even cutely-confused, read Monika, people. I am fortunate to have eaten many of these Delhiite friends' lunch boxes. I have felt a resonance with the dark humour of Ajay, Hareesh, and Prateekshit. I have never felt time passing faster than when I am talking with them. Aastha, Hemank, Janhavi, Aishwariya, Mahek, Priyanka, Gaurav, Soumyadeep, Rohit, Talha, and Praneet have not only been my collaborators, but have helped me broaden my horizons too.

I would like to thank Tata Consultancy Services and Department of Electronics and Information Technology, Government of India for their funding and support. I would also like to thank Vivek ji, Sheetu ji, and Prosenjit ji from the Administration who have helped keep things smoother.

During my B.E. final year when my father passed away, I had felt a need to take up a job right after undergraduate, and to leave my dream of higher studies as a dream. However, it was my mother, brother, and bhabhi who consoled me, and gave me a confidence to revert to my earlier goal of pursuing higher studies. Had they not done so, quite likely, I would have been a software developer all this time. My father used to advise to be a good citizen and to give back to society. I hope my training and education eventually helps me do so.

Finally, I would like to thank love of my life, my wife, my shareek-e-hayat, Chandani for all the support she has provided me. To get married to a PhD pursuing guy is quite an extraordinary decision in itself, at least in Indian scenario. She did it, she believed in me, she brought me up from my lows, she became my strength and continues to be so. Without her support and understanding, things could have fallen apart in more ways than one. If I am able to get a sense of things falling into place, quite certainly, it is because of her.

Dedicated To My Family

(Yes, it includes you too, my little one)

Pushing Boundaries of Face Recognition: Adversary, Heterogeneity, and Scale

by

Tejas Indulal Dhamecha

Abstract

Due to the unconstrained nature of data capture and non-cooperative subjects, automatic face recognition is still a research challenge for application scenarios such as law enforcement. We observe that challenges of face recognition are broadly rooted into two facets: (1) the non-ideal and possibly adversarial face image samples and (2) the large size and incremental/streaming availability of data. The first facet encompasses various challenges such as intentional or unintentional obfuscation of identity, attempts for spoofing system, user non-cooperation, and large intra-subject variations for heterogeneous face recognition. The second facet caters to challenges arising due to application scenarios such as repeat offender identification and surveillance where the data is either large scale or available incrementally. Along with advancing the face recognition research by addressing the challenges arising from both the aforementioned facets, this dissertation also contributes to the pattern classification research by abstracting the research problems at the classifier level and proposing feature independent solutions to some of the problems.

The first contribution addresses the challenge of face obfuscation due to usage of disguise accessories. We collect and benchmark IIIT In and Beyond Visible Spectrum Face Dataset (I^2BVSD) pertaining to 75 subjects, which has various types of disguises applied on different individuals. It has become one of the most used disguise face dataset in the research community. Since disguised facial regions can lead to erroneous identity prediction, a texture based algorithm is designed to differentiate between biometric and non-biometric facial patches. The proposed approach is embedded with local face recognition algorithm to address the challenge of disguise variations. The approach is further enhanced with the use of thermal spectrum imaging. As the second contribution, the dissertation addresses the challenge of heterogeneous face matching scenarios, such as matching a sketch against a mugshot dataset of digital photographs, cross-spectrum, and cross-resolution matching, that arise in a wide range of law enforcement scenarios. Heterogeneous Discriminant Analysis (HDA) is designed to encode multi-view heterogeneity in the classifier to obtain a projection space more suitable for matching. Further, to extend the proposed technique for nonlinear projections, formulation of kernel HDA is proposed. Focusing on application such as identification of repeat offenders, as the third contribution, we develop an approach to efficiently update the face recognition engine to incorporate incremental training data. The proposed Incremental Semi-Supervised Discriminant Analysis (ISSDA) provides mechanism to efficiently, in terms of accuracy and training time, update the discriminatory projection directions. The proposed approach capitalizes on offline unlabeled face image data, which is inexpensive to obtain and generally available in abundance. The fourth contribution of this dissertation is focused on designing

a face recognition classifier that can be efficiently learned from very large batches of training data. The proposed approach, termed as Subclass Reduced Set Support Vector Machine (SRS-SVM), utilizes the subclass structure of training data to effectively estimate the candidate support vector set. This candidate support vector set facilitates learning of nonlinear Support Vector Machine from large-scale face data in less computation time.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1 Introduction **1**

1.1 Unconstrained Environments and Scale: Two Challenges of Face Recognition . . . 5

1.1.1 Unconstrained Environment: Adversary and Heterogeneity 7

1.1.2 Scale of Data: Incremental and Large Scale Training 10

1.2 Role of Classifiers in Face Recognition Pipeline 13

1.3 Research Objectives and Contributions 15

2 Recognizing Disguised Faces **19**

2.1 Material and Methods 23

2.1.1 Ethics 23

2.1.2 Disguise Face Database 25

2.1.3 Participants for Human Evaluation 26

2.1.4 Stimuli, Design and Procedure 27

2.1.5 Observations from Human Evaluation 30

2.2 Anāvṛta: Proposed Face Recognition Approach 36

2.2.1 Patch Classification 38

2.2.2 Patch based Face Recognition 40

2.2.3 Results of the Proposed Algorithm 40

2.2.4 Comparison of Human Responses with Automated Algorithms 45

2.3 Summary 47

3 Heterogeneous Discriminant Analysis for Cross-View Face Recognition **49**

3.1 Heterogeneous Discriminant Analysis 54

3.1.1 Adaptation of scatter matrices . . . 56
3.1.2 Visualization . . . 63
3.2 Proposed Face Recognition Approach . . . 64
3.3 Experimental Evaluation . . . 66
3.3.1 Cross Spectral Face Matching: Visible to NIR Images . . . 67
3.3.2 Cross Resolution Face Matching . . . 72
3.3.3 Digital Photo to Composite Sketch Face matching . . . 74
3.4 Comparison with Related Approaches . . . 76
3.5 Summary . . . 78

4 Incremental Semi-supervised Discriminant Analysis for Face Recognition 79
4.1 Incremental Semi-supervised Discriminant Analysis . . . 83
4.1.1 Semi-Supervised Discriminant Analysis . . . 84
4.1.2 Incremental Learning . . . 86
4.1.3 Time Complexity . . . 91
4.2 Experiments and Results . . . 91
4.2.1 Database, Experiment Design, and Protocols . . . 92
4.2.2 Experiment 1: Comparative Evaluation . . . 95
4.2.3 Experiment 2: Effect of Manifold Regularization . . . 101
4.2.4 Experiment 3: Effect of Size of Unlabeled Set . . . 102
4.2.5 Experiment 4: Incremental Addition of Classes . . . 106
4.3 Summary . . . 107

5 Large-scale Face Recognition by Leveraging Subclasses in Kernel SVM 109
5.1 Preliminaries of SVM . . . 115
5.2 Reduced Set and Variants . . . 117
5.3 Proposed Subclass Reduced Set SVM . . . 118
5.3.1 Estimating Minimal Representative Reduced Set . . . 119
5.3.2 Hierarchical Subclass Reduced Set SVM (HSRS-SVM) . . . 124
5.4 Datasets and Protocols . . . 127
5.5 Experiments on Synthetic Datasets . . . 131

5.5.1 Visualization of Each Step 134
5.5.2 Quantitative Analysis 135
5.6 Experiments on Real-world Datasets 137
5.6.1 Comparative Analysis 138
5.6.2 Training Time of Individual Stage 139
5.6.3 Effectiveness of MRRS Estimation Approach 140
5.6.4 Effect of h (Number of Subclasses) and μ (Number of Children) Parameters in Hierarchical SRS-SVM 143
5.6.5 HSRS-SVM with Deep Learning Features for Face Recognition 145
5.7 Summary 147

6 Conclusion and Future Research Directions **149**

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

1-1 Illustrating the procedure for acquiring face, fingerprint, and iris samples. 2

1-2 CCTV cameras can be installed for the purpose of surveillance. The images show CCTV cameras installed at entrance of a premise to keep record of the visitors. . . 3

1-3 Broad overview of various challenge of face recognition in different use-cases. Traditionally, the focus has been on the covariates of pose, illumination, and expression. For pushing face recognition research further, issues pertaining to adversarial user behaviour, imaging heterogeneity needs to be addressed to mitigate the effects of unconstrained environments. Efficient training from large-scale and incremental sources is important to address the effects of increased operational scope. This dissertation focuses on addressing challenges associated with adversary, heterogeneity, and scale for face recognition. 5

1-4 Illustrating some of the challenges of face recognition in unconstrained environment. 6

1-5 Examples of heterogeneous face recognition scenarios. Images in (a) show heterogeneity due to spectrum difference, (b) illustrate heterogeneity due to resolution differences, (c) show sample digital photos and their corresponding composite sketches respectively. 7

1-6 Illustrating the need to update the recognition engine with incrementally available face samples as the appearance may change. The image shows face images of a subject acquired over several years. Updating such (intra- and inter-) class characteristics in the classifier models is necessary. 12

1-7 A typical face recognition pipeline. 13

2-1 Illustrating the effect of disguise accessories on inter-class and intra-class variations. . . . 19
2-2 Sample questionnaire. . . . 24
2-3 Sample images from the ID V1 database. . . . 25
2-4 Distribution of genuine and impostor pairs in questionnaires. . . . 29
2-5 Effect of disguising individual facials parts and their combinations. . . . 35
2-6 Illustrating the steps involved in the proposed face recognition framework. . . . 37
2-7 The distribution of biometric patches in the training and test sets. . . . 41
2-8 ROC curves for patch classification. . . . 42
2-9 The results of the proposed face recognition framework using LBP descriptor. . . . 43
2-10 Performance of disguised face recognition by humans, with respect to familiarity and ethnicity factors. . . . 46

3-1 Examples of heterogeneous face recognition scenarios. Top row (a) & (b) shows heterogeneity due to spectrum difference. The middle row (c)-(f) illustrates heterogeneity due to resolution differences. (The images of different resolution are stretched to common sizes.) The bottom row shows (g)-(h) shows photo and composite sketches of the two subjects. . . . 50
3-2 An example illustrating heterogeneous and homogeneous matching. Here, two views pertaining to spectrums (VIS and NIR) are shown. The solid lines represent comparisons corresponding to heterogeneous matching. . . . 51
3-3 HDA vs LDA: Toy example . . . 56
3-4 Graphical interpretation of the proposed HDA. . . . 58
3-5 Steps involved in the face recognition pipeline with the proposed HDA and KHDA. 65
3-6 ROC curves on the CASIA NIR-VIS-2.0 database [146]. . . . 70
3-7 CMC curves for cross-resolution matching (Probe: 48×48, 32×32, 24×24, 16×16, Gallery: 72×72) on the CMU-MultiPIE database [61]. . . . 73
3-8 Performance improvement due to HDA and KHDA with LCSSE features on the CMU Multi-PIE database. Size of gallery images is 48×48, whereas probe sizes are 32×32, 24×24, 16×16. . . . 74

3-9 CMC curve for composite sketch to digital photo matching on the e-PRIP composite sketch dataset [161], [162]. 75

4-1 Traditional incremental discriminant approaches, such as Kim *et al.* [207], [208] and Lamba *et al.* [211], update between-class and overall variability. New eigenmodels of $\mathbf{S}_B$ and $\mathbf{S}_T$ are learned from incremental batch, which are merged with corresponding existing eigenmodels. Discriminating components $\mathbf{V}$ are obtained from merged eigenmodels of $\mathbf{S}_B$ and $\mathbf{S}_T$. 80

4-2 The proposed approach incrementally learns the between-class variability and uses unlabeled data to learn the overall variability. Eigenmodel of $\mathbf{S}_B$ is learned from incremental batch and merged with the existing eigenmodel. New discriminating components $\mathbf{V}$ are obtained using updated eigenmodel of $\mathbf{S}_B$ and offline estimated eigenmodel of $\mathbf{S}_T$. 85

4-3 Block diagram of the evaluation protocol for face recognition experiments. At first, the model is learned using initial training samples and unlabeled data. With each incremental training batch the existing model is updated to obtain a new model. . . 89

4-4 Sample images from the (a) CMU-PIE dataset [55] (b) visible spectrum and (c) NIR images from VIS-NIR-2.0 dataset [228], and (d) CMU-MultiPIE dataset [61]. 93

4-5 Rank-1 identification accuracy for sub-experiments pertaining to CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE face dataset. The graph representation of the accuracy helps understand the general trend among approaches. 99

4-6 CMC curves of the proposed approach for (a) CMU-PIE, (b) NIR-VIS-2.0 (VIS only), (c) NIR-VIS-2.0 (NIR only), and (d) CMU-MultiPIE datasets. Consistently, the incremental update improves the identification performance. 100

5-1 Illustrating four categories of approaches designed for scalable SVM learning. . . 111

5-2 Abstract illustration explaining the core concept of the proposed approach, Subclass Reduced Set SVM. Approaches, such as SRS-SVM, that fall under the categorizations of the subset based and piece-wise linear approaches, operate on this basic intuition. 118

5-3 Traditionally SVM solver is applied on the complete training set. The proposed SRS-SVM operates in two stages: estimating MRRS and applying SVM solver on the obtained reduced set. For a detailed illustration of MRRS estimation block, refer Figure 5-4. 119

5-4 Block diagram of MRRS estimation procedure of the proposed Subclass Reduced Set SVM. Each class is divided into h subclasses. Each subclass of $+1$ class is paired with each subclass of -1 class, thus resulting in a total of h^2 subclass-pairs. Support vectors from each subclass-pair are retained as the candidate global support vectors. They are combined either using union operator (in SRS-SVM) or using a further hierarchical aggregation (in Hierarchical SRS-SVM). 120

5-5 Illustrating the applicability of subclass structure in modeling decision boundary for Dog vs Cat classification problem. Out of a vast variety of dog and cat breeds, there are only limited breeds (subclasses) that contribute to the decision boundary. Further, different decision boundaries between breed-pairs of dogs and cats can be seen as constituents for the overall decision boundary. 121

5-6 Graphical illustration of the proposed Hierarchical Subclass Reduced Set SVM (HSRS-SVM). 126

5-7 Illustration the synthetic datasets used for performance evaluation (best viewed in color). 128

5-8 Samples of the real world databases used for performance evaluation: (a) animal and non-animal class images from `CIFAR-10` [243], [271] and (b) face and non-face images from face detection dataset of Pascal Large Scale Learning Challenge [273] 130

5-9 Visualization of proposed approach on the `XOR` dataset. Training on whole dataset ($n = 800, h = 2$) LibSVM takes 3.46 seconds; whereas the proposed SRS-SVM obtains similar decision boundary in 0.25 seconds. See Algorithm 2 to relate the mathematical formulation of the individual steps. 132

5-10 Illustrating the processing of the proposed SRS-SVM on the Shooting Range dataset. Training on the whole dataset ($n = 4{,}500$) LibSVM takes 93 seconds; whereas the proposed SRS-SVM obtains similar decision boundary in 50 seconds. 133

5-11 Comparative illustration of the decision boundaries obtained by LibSVM and by the proposed SRS-SVM approach ($h = 5$). 134

5-12 Comparing training time on three synthetic datasets: two concentric circles (2CC), three concentric circles (3CC), and XOR. A varying number of samples are generated for each of the datasets. The training time is shown on the logarithmic scale. As the number of training instances increases, the training time of LibSVM increases rapidly whereas, the proposed SRS-SVM has a significantly lower rate of increase in training time. 136

5-13 Comparing training time on SR (Shooting Range) dataset. Different number of samples are generated from the dataset and training set size vs training time plots is shown for different dataset sizes with number of subclasses (h) as 5, 15, and 20. Consistently, SRS-SVM takes less training time compared to LibSVM. As the parameter h is increased, the training time is observed to reduce significantly on the logarithmic scale. 137

5-14 Sample images for Labeled Faces in the Wild (LFW) dataset. The classification task involves verifying if the identity of persons in two images is same (match pair) or not (non-match pair). 145

5-15 ROC curves on restricted protocol of LFW dataset. 147

6-1 4Vs of face recognition: Classification of face recognition challenges for next generation recognition systems. 151

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

1.1 Application scenarios as function of user behaviour and environment settings. . . . 9

1.2 Summary of various public face datasets. Notice the increase in the size (number of subjects and samples) of the datasets in recent years. 11

2.1 Literature review: Disguise face recognition 22

2.2 Age and gender distribution of participants in the four sets. 28

2.3 Summary of human performance. 30

2.4 p-values of statistical tests to understand the effect of each factor. 32

2.5 Confusion matrix for comparing the consistency of Set FS-I and Set FS-II. 34

2.6 Verification results obtained from automated algorithms. 44

3.1 Summary of selected research papers on heterogeneous face matching. VS=viewed sketch, FS=forensic sketch, CR=cross resolution, HR=high resolution, LR=low resolution, TH=thermal, VIS=visible, and NIR=near infrared. 53

3.2 Analyzing the overlap of projection distributions in Fig. 3-3b and Fig. 3-3c. 64

3.3 Summary of the datasets. 65

3.4 Rank-1 identification accuracies for visible to near infrared face matching on the CASIA NIR-VIS-2.0 database [146]. The experiments are performed by varying the feature extractors, classification models, and distance metrics. 67

3.5 Face recognition performance of the proposed and some existing algorithms for VIS to NIR face matching on CASIA NIR-VIS-2.0 dataset. †represents value obtained from ROC curve reported in the corresponding paper. *represents the results reported in [121], [147]. 69

3.6 Rank-1 identification accuracy of the proposed HDA, KHDA and existing algorithms on CMU-MultiPIE database with different gallery and probe image sizes. Two top performing approaches are highlighted in each cross-resolution setting. . 71

3.7 Rank-10 Identification accuracy for composite sketch to photo matching. The results marked with $*$ are reported by Mittal *et al.* [162]. 76

4.1 Literature review of the related research pertaining to incremental learning and semi-supervised learning algorithms related to discriminant analysis. 81

4.2 Computational complexity analysis. M and D represent the number of samples and feature dimensionality. $d_{T,i}$ and $d_{B,i}$ is the number of components preserved in eigenmodels of total and between-class scatter matrices of i^{th} batch. M_u is the number of samples in unlabeled set and k is the neighborhood parameter of learning graph laplacian. 92

4.3 Experimental protocols 94

4.4 Rank-1 identification accuracy (mean$\pm$std-dev %) and computation time for sub-experiments pertaining to the CMU-PIE, and NIR-VIS-2.0 (VIS spectrum). Initial batch consists of one image per subject and each incremental batch consists of one new image per subject. Therefore, there are n images in initial and each incremental batch, where n is the number of subjects in the test split. *Algorithms are not incremental, therefore batch mode training results are reported. 97

4.5 Rank-1 identification accuracy (mean$\pm$std-dev %) and computation time for sub-experiments pertaining to the NIR-VIS-2.0 (NIR spectrum) and CMU-MultiPIE face datasets. Initial batch consists of one image per subject and each incremental batch consists of one new image per subject. Therefore, there are n images in initial and each incremental batch, where n is the number of subjects in the test split. *Algorithms are not incremental, therefore batch mode training results are reported. 98

4.6 Confusion matrix for comparing the performance of SSDA and ISSDA. ✓ and ✗ represent the percentage of correctly classified and misclassified samples respectively. 102

4.7 Rank-1 identification accuracy (mean±std-dev %) on the CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE dataset without using manifold regularization. The experiment design and protocol is same as Experiment 1 except that $\beta_1 = 0$ is set. For easier comparison, the results obtained with manifold regularization are shown reported within brackets. The smaller the dataset the more noticeable is the performance drop when manifold regularization is not used. 103

4.8 Rank-1 identification accuracy (mean±std-dev %) on the CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE datasets with varying sizes of unlabeled set. It can be observed that moving from left to right in a row (increasing the size of unlabeled set), and top to bottom (updating the model with incremental batches) yields better accuracy. 104

4.9 Rank-1 identification accuracy (mean±std-dev %) on the CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE datasets by incrementally adding new subjects. The accuracy difference when incrementally adding new subjects (classes) is similar to that of incremental addition of new images of existing subjects as reported in Table 4.4 and 4.5. 105

5.1 Training time as a function of the number of training instances for a synthetic two-dimensional dataset `two concentric circles` (2CC). 109

5.2 The effect of the number of subclasses on the size of estimated MRRS. 125

5.3 Details pertaining to the real-world datasets used in the evaluation and their corresponding hyperparameters. (d is feature dimensionality, h is number of subclasses, C is misclassification cost, and γ is radial basis function kernel parameter) 131

5.4 Results of the proposed HSRS-SVM in comparison to other related approaches in terms of training time (in seconds) and classification accuracy (in percentage). . . . 138

5.5 Stage-wise training time of the proposed subclass reduced set based SVM approach. Time is reported in seconds. The figures in the parenthesis represent the fraction of total training time consumed in percentage. Level 2 is the root level as $\mu = h^2$. 140

5.6 Numerical analysis of the precision and recall of the estimated minimal reduced representative set ($\hat{T}_{MRRS}$) and the final support vector set (T_{rSV}) obtained using proposed HSRS-SVM approach with respect to the support vector set (T_{SV}) of the traditional solver (LibSVM). 142

5.7 Effect of varying number of subclasses (h) and number of children (μ) on the training time and classification accuracy of the proposed HSRS-SVM on the `adult` dataset. The training time is reported in seconds. The figures within parenthesis represent the classification accuracy. 143

5.8 Verification accuracy of utilizing LCSSE features with HSRS-SVM ($h = 5, \mu = 25$) and LibSVM in comparison to state-of-the-art approaches. 146

THIS PAGE INTENTIONALLY LEFT BLANK

Abbreviations

ACC Accuracy.

CCTV Closed-circuit Television.

CMC Cumulative Match Characteristic.

COTS Commercial Off-The-Shelf.

CS Composite Sketch.

EER Equal Error Rate.

FA False Accepts.

FAR False Accept Rate.

FR False Rejects.

FRR Falser Reject Rate.

GA Genuine Accepts.

GAR Genuine Accept Rate.

GR Genuine Rejects.

HDA Heterogeneous Discriminant Analysis.

HR High Resolution.

I^2BVSD IIIT-D In and Beyond Visible Spectrum Face Dataset.

ISSDA Incremental Semi-Supervised Discriminant Analysis.

LR Low Resolution.

NIR Near-Infrared (spectrum).

PIE Pose, Illumination and Expression.

RGB Red, Green and Blue.

ROC Receiver Operating Characteristic.

SRS-SVM Subclass Reduced Set Support Vector Machine.

VIS Visible (spectrum).

Research Dissemination

Journals

Published

[1] **T. I. Dhamecha**, R. Singh, and M. Vatsa, "On incremental semi-supervised discriminant analysis," *Pattern Recognition*, vol. 52, pp. 135 –147, 2016, **[Impact Factor: 5.482]**.

[2] **T. I. Dhamecha**, R. Singh, M. Vatsa, and A. Kumar, "Recognizing disguised faces: Human and machine evaluation," *PLoS ONE*, vol. 9, no. 7, e99212, 2014, **[Impact Factor:2.806]** [Rank# 25 Scientific Journal with h5-index of 171 as per Google Scholar].

Under Review

[1] **T. I. Dhamecha**, R. Singh, and M. Vatsa, "Heterogeneous discriminant analysis for cross-view face recognition," 2017.

[2] **T. I. Dhamecha**, A. Noore, R. Singh, and M. Vatsa, "Between-subclass piecewise linear solutions for learning large scale kernel SVM," 2017.

[3] **T. I. Dhamecha**, M. Shah, P. Verma, M. Vatsa, and R. Singh, "CrowdFaceDB: Database and benchmarking for face verification in crowd," *Pattern Recognition Letters*, 2017, (Minor Revision Received) **[Impact Factor: 1.995]**.

Book Chapters

[1] J. Agrawal*, A. Pant*, **T. I. Dhamecha***, R. Singh, and M. Vatsa, "Understanding thermal face detection: Challenges and evaluation," in *Face Recognition Across the Electromagnetic Spectrum*, Springer, 2015, pp. 139–163.

[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, 2015, pp. 279–297.

Peer Reviewed Conference Articles

[1] **T. I. Dhamecha***, P. Sharma*, R. Singh, and M. Vatsa, "Discriminative facetopics for face recognition via latent dirichlet allocation," in *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[2] S. Ghosh*, **T. I. Dhamecha***, R. Kesari, R. Singh, and M. Vatsa, "Feature and keypoint selection for visible to near-infrared face matching," in *IEEE International Conference on Biometrics: Theory, Applications & Systems*, 2015.

[3] **T. I. Dhamecha***, P. Verma*, M. Shah*, R. Singh, and M. Vatsa, "Annotated crowd video face database," in *IAPR International Conference on Biometrics*, 2015, pp. 106–112.

[4] **T. I. Dhamecha***, P. Sharma*, R. Singh, and M. Vatsa, "On effectiveness of histogram of oriented gradient features for visible to near infrared face matching," in *International Conference on Pattern Recognition*, 2014, pp. 1788–1793.

[5] **T. I. Dhamecha**, A. Nigam, R. Singh, and M. Vatsa, "Disguise detection and face recognition in visible and thermal spectrums," in *IAPR International Conference on Biometrics*, 2013, pp. 1–8.

[6] S. Bharadwaj, **T. I. Dhamecha**, M. Vatsa, and R. Singh, "Computationally efficient face spoofing detection with motion magnification," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 105–110.

[7] H. Lamba, **T. I. Dhamecha**, M. Vatsa, and R. Singh, "Incremental subclass discriminant analysis: A case study in face recognition," in *IEEE International Conference on Image Processing*, 2012, pp. 593–596.

[8] **T. I. Dhamecha**, A. Sankaran, R. Singh, and M. Vatsa, "Is gender classification across ethnicity feasible using discriminant functions?" In *IEEE/IAPR International Joint Conference on Biometrics*, 2011, pp. 1–7.

* These authors contributed equally.

Chapter 1

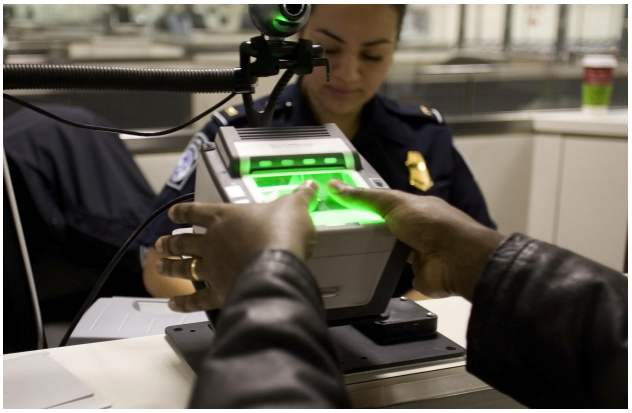
Introduction

Biometrics is the field of study that deals with identifying humans based on their physiological and behavioural traits [1]. The various traits/modalities that are explored for biometric authentication include physiological traits such as face [2], fingerprint [3], iris [4], retina [5], palmprint [6], knuckle print [7], hand geometry [6], and ear [8] and behavioural traits such as gait [9], signature [10], and keystroke dynamics [11]. Among these, face, fingerprint, and iris are, arguably, the most prominent, popular, and widely implemented modalities. Figure 1-1 illustrates procedures for collecting face, fingerprint, and iris samples. Fingerprint and iris modalities require specialized hardware and expect relatively higher degree of user cooperation, whereas, a face image can be collected by placing a commercially available imaging device within a distance of up to few meters of the subject. Further, it is possible to capture face images without restricting or causing discomfort to the user, by utilizing devices such as CCTV camera, as shown Figure 1-2. This *non-intrusiveness* is an important characteristic in various application scenarios such as law enforcement and surveillance where the user may be freely moving and is not necessarily cooperative. Thus, for certain application scenarios, we find face to be a relatively more suitable modality.

The biometrics research community has actively focused on investigating automatic methods to recognize faces[1] for several decades [2], [12]. Some important face recognition approaches, in chronological order, are Geometric approach [13], EigenFaces [14], Local Feature Analysis [15], FisherFaces [16], Elastic Bunch Graph Matching (EBGM) [17], Gabor features [18], [19], Bayesian learning based approach [20], Local Binary Patterns (LBP) [21], Scale Invariant Feature

[1] In this dissertation the term machine and computer are interchangeably used.

(a) Face



(b) Fingerprint

(c) Iris

Figure 1-1: Illustrating the procedure for acquiring face, fingerprint, and iris samples. (a) Face image is acquired using front facing camera of a handheld mobile device. (b) Optical sensor based fingerprint acquisition device used as part of US-VISIT (Visitor and Immigrant Status Indicator Technology) program. (c) Iris image is captured using a specialized handheld device.
Image Sources: goo.gl/htrLxZ, https://goo.gl/tmH5vE, https://goo.gl/7riXQC

Figure 1-2: CCTV cameras can be installed for the purpose of surveillance. The images show CCTV cameras installed at entrance of a premise to keep record of the visitors.

Transform (SIFT) [22], [23], Dictionary-learning based approaches [24]–[26], Sparse Representation Classifier (SRC) [27], Joint Bayesian learning [28], Fisher vector faces [29], and Deep Learning based algorithms [30]–[33].

State-of-the-art on various benchmark datasets is reported by deep learning based approaches. These approaches are widely based on Convolutional Neural Networks (CNN). Although the core idea of utilizing CNNs for face recognition existed for about two decade [34], the major impediments to leverage it fully were mostly rooted in limited data and computation power. With the advancements in the parallel processing hardwares, e.g. general purpose graphics processing units, and neural networks training algorithms, it has become possible to achieve impressive results using CNN based approaches. Invariably, almost all the deep learning based approach involves learning about hundred million parameters of the underlying neural network architecture. Some of the top performing approaches include DeepFace [35], FaceNet [36], DeepID [30], [31]. It should be noted that these approaches also involve state-of-the-art pre-processing techniques and metric/classifier learning. Broadly, the CNN learns primitive to complex features in the subsequent layers. It is observed that in the first few layers, CNN learns features similar to edges and hand-crafted filters (e.g. Gabor). The availability of large labelled data, massive computing power, advances in learning algorithms have brought machine face recognition on some benchmarks on par with humans.

The earlier research majorly focused on addressing each covariate, such as pose, illumination, and expression (PIE), individually. The research succeeded in demonstrating the potential of face recognition for various well-controlled scenarios. In due course of time, researcher have been broadening the scope of face recognition to increasingly uncontrolled scenarios. For example, law enforcement related application scenarios such as surveillance, assume very limited control over user or environment. A broad view of the various challenges of using face recognition in different application scenarios is illustrated in Figure 1-3. Face recognition algorithms have achieved impressive accuracy in controlled environments [37], [38], i.e. frontal pose, moderate expression, and controlled illumination.

The advancements have led to adoption of face recognition in various e-governance and commercial scenarios such as e-passport, access control, and attendance systems. These application scenarios provide a significant control over the imaging environment and the users. Some examples of face recognition technology in real-world scenarios include Australia's automated border processing system[2], face recognition based user authentication on mobile phones (FaceLock) and laptops (Windows 10) [39], face clustering and tagging in *Picasa* and *iPhoto*, and face-tag recommendation functionality in *Facebook*. Similarly, *UIDAI* (Unique Identification Authority of India)[3] and *US-VISIT* (Visitor and Immigrant Status Indicator Technology)[4] programs also collect face image along with other biometric samples. However, in all these cases it is likely that the face images are captured with user cooperation in controlled environment. Thus, it is broadly accepted that state-of-the-art face recognition systems have matured to be useful at least within the constraints of controlled environment and user cooperation [37], [38]. In recent years, unconstrained face recognition, has also attained significant advances especially by utilizing deep learning based approaches [30]–[33].

As illustrated in Figure. 1-3, we believe that various challenges of face recognition can be brought under a broad categorization of *unconstrained environment* and scale. These challenges can be abstracted and the proposed solutions can have broader applicability. While these challenges can be addressed at various stages of face recognition, addressing some of these challenges at classifier level is more suitable and/or effective. Therefore, in this dissertation, we focus on devising

[2] https://goo.gl/EcLJxJ, last retrieved: 15 Jun 2017
[3] https://uidai.gov.in/beta/, last retrieved: 15 Jan 2017
[4] https://www.dhs.gov/obim, last retrieved: 15 Jan 2017

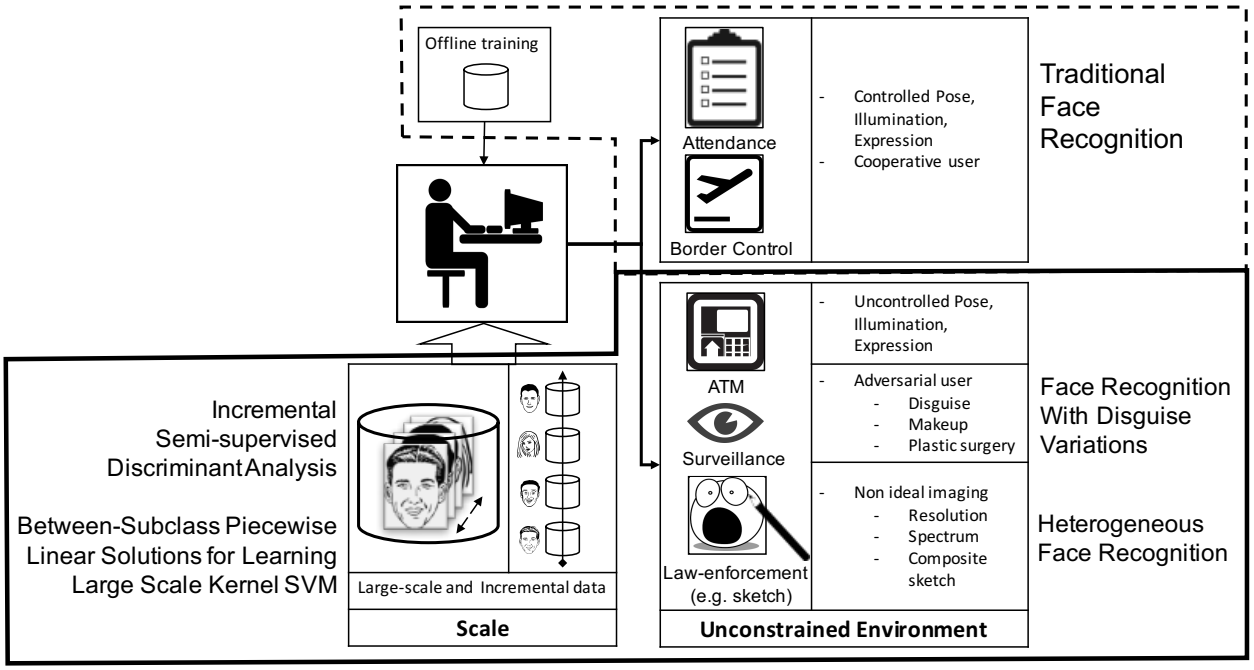


Figure 1-3: Broad overview of various challenge of face recognition in different use-cases. Traditionally, the focus has been on the covariates of pose, illumination, and expression. For pushing face recognition research further, issues pertaining to adversarial user behaviour, imaging heterogeneity needs to be addressed to mitigate the effects of unconstrained environments. Efficient training from large-scale and incremental sources is important to address the effects of increased operational scope. This dissertation focuses on addressing challenges associated with adversary, heterogeneity, and scale for face recognition.

classifier level solutions for addressing challenges associated with unconstrained environment and scale. Further details regarding challenges and the nature of their solution are discussed in the following sections.

1.1 Unconstrained Environments and Scale: Two Challenges of Face Recognition

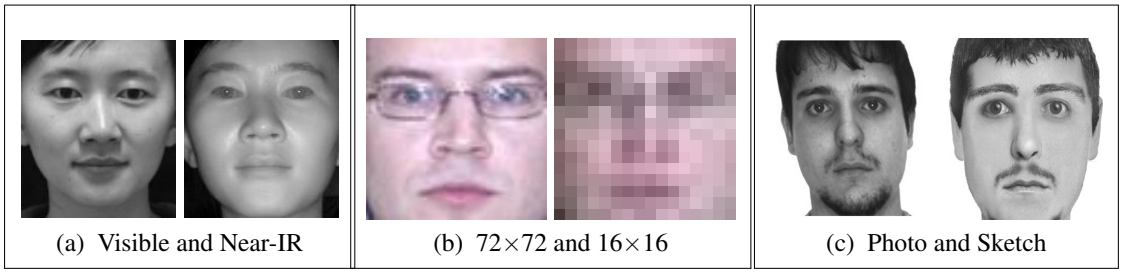
As illustrated in Figure 1-3, application scenarios such as law enforcement and surveillance present novel challenges. The roots of important challenges can be broadly traced to two aspects: 1) unconstrained environment and 2) scale of data.

(a) Controlled Photo (b) Unconstrained Photo

(c) Variations in face appearance due to usage of accessories. An adversarial subject can elude from automatic recognition systems by using such facial accessories.

(d) Image of a crowd captured at a distance resulting in low resolution of individual faces

Figure 1-4: Illustrating some of the challenges of face recognition in unconstrained environment.



(a) Visible and Near-IR (b) 72×72 and 16×16 (c) Photo and Sketch

Figure 1-5: Examples of heterogeneous face recognition scenarios. Images in (a) show heterogeneity due to spectrum difference, (b) illustrate heterogeneity due to resolution differences, (c) show sample digital photos and their corresponding composite sketches respectively.

1.1.1 Unconstrained Environment: Adversary and Heterogeneity

Face images captured under different environmental conditions can vary significantly. For example, in law enforcement scenarios, the users/subjects are often the suspects which may or may not be cooperative in allowing face image acquisition. Further, the imaging environment (e.g. public places, outdoors, nighttime) itself is often uncontrolled. Thus, a significant portion of variations is rooted in degree of user/subject cooperation and imaging environment. In scenarios where user cooperation can be expected (e.g. visa application, attendance systems), the variations due to pose, illumination, and expression can be avoided. Similarly, ability to control the imaging environment (e.g. studio) can also reduce the effects of these covariates. However, in application scenarios, such as law enforcement, the control over user cooperation and imaging environment is limited. Table 1.1 briefly summarizes the expected operating settings for law enforcement application scenarios, i.e. imaging environment may be controlled or uncontrolled and the user is not necessarily cooperative. The images in Figure 1-4, although not captured in actual law enforcement scenario, illustrate the variations in face images of the same subject depending on acquisition settings. Perceivably, the images captured with limited-or-no user cooperation can be challenging to recognize. Notice the difference between the passport photo (Figure 1-4a) acquired under controlled environment and the so-called wild photo (Figure 1-4b) captured without control over PIE (pose, illumination, and expression) and user cooperation. As shown in Figure 1-4c, there is a significant difference in the appearance among images due to utilization of facial accessories. Such attempts to appear different or to appear similar to other subjects can be categorized into adversarial user behaviour. Figure 1-4d shows an example of non-ideal image acquisition that results into low resolution and

occlusion along with illumination and expression variations.

The query face images in the application scenario of law enforcement are likely to be acquired in *unconstrained environment without user cooperation*. For example, in case of surveillance in public places, such as metro stations, people may be moving, parts of some faces may be occluded, and there can be non-uniform illumination and various imaging artifacts present in the acquired sample. Such challenges are likely to be present in real-world unconstrained face recognition scenarios. Unconstrained nature also leads to non-ideal image acquisition; because of which faces may be captured at low resolution, at long distance stand-off, and using sensors operating in different spectrums of light. Further, in suspect identification scenario, the input/query face may be a hand-drawn sketch or a computer generated composite sketch. These scenarios stem challenges such as sensor interoperability, cross-spectrum, photo-to-sketch, cross-resolution, and long stand-off face recognition. Figure 1-5 shows such examples of heterogeneity introduced due to non-idea imaging scenarios. Further, the subjects may even exhibit adversarial behavior with the help of masks, facial accessories, and surgical modifications [40]–[43]. Thus, face recognition in unconstrained environments involves challenges which are not encountered in typical controlled environment scenarios.

Various face recognition challenges and NIST (National Institute of Standards and Technology) evaluations have quantitatively emphasized that performance of state-of-the-art systems have been significantly higher for matching face images acquired in controlled environment (e.g. mugshot) compared to other scenarios (e.g. poor quality webcam images) [37], [38], [44]. These evaluations also provide insights into how much the face recognition approaches have evolved in addressing various challenges. Following is the summary of some important benchmark evaluations to provide insights into the need for uncontrolled face recognition.

- **MBE 2010: Visa Application vs Law Enforcement Mugshot Images:** Multi-Biometric Evaluation, 2010 [37] reports that, overall, face verification performance has improved for good quality images such as the ones captured during visa processing. Between 2002 and 2010, the improvements in face recognition engines have led to reduction in FRR from 20% to 0.3% at 0.1% FAR on good quality images. In identification scenario, the accuracy of the best matcher on law enforcement mugshots dataset is 3% lower than that on the Visa Application dataset. Further, the effect of scale (population size) on the accuracy is also

Table 1.1: Application scenarios as function of user behaviour and environment settings.

		Environment	
		Controlled	Uncontrolled
User	Cooperative	VA, LE	-
	May not be cooperative	LE	
	Adversarial		

VA = Visa Application, LE = Law Enforcement

observed. The report suggests that there is an approximate dependence of accuracy on log of the population size. Quantitatively, for the populations sizes of N=10K, 80K, 320K, and 1.6M, the rank-1 accuracy is observed as 96.9%, 95.3%, 93.6%, and 92.3% respectively.

- **GBU 2011: Quantitative Effects of Covariates:** Good, Bad, and Ugly face recognition challenge [45] studies the recognition performance under fixed settings of pose, aging, and image acquisition. It shows that due to *the way faces are represented in the image* there can be sets of image-pairs with very high, moderate, or poor classification accuracy. These three kinds of sets are termed as good, bad, and ugly. It reports that at FAR of 0.1%, the baseline algorithm achieves FRR of 2%, 20%, and 85% on good, bad, and ugly sets, respectively. Since the protocol controls the pose, aging, and sensor variations, the performance difference is largely attributed to the variations in environments (studio settings, hallways, atria, or outdoors), illuminations, and expressions.

- **FRVT 2014:** The Face Recognition Vendor Test 2014 [38] further provides insights into the challenges of face recognition in law enforcement applications. One of the experiments evaluates the performance of matching mugshot images against webcam images which are acquired in relatively uncontrolled environment. As the webcam based image acquisition imposes relatively less constraints on the subject, the webcam-to-webcam (93.4%) image matching yields poorer performance than mugshot-to-mugshot (97.5%). Further, the cross-sensor interoperability challenge is evident in the observation that mugshot-to-webcam matching achieves only 89.6% accurate face recognition.

Recently, NIST has launched three major challenges focused around unconstrained face recogni-

tion; namely IJB-A [46], IJB-B [47], and Face Recognition Prize Challenge[5]. These challenges have provided platforms and testbeds for evaluating face recognition performances in unconstrained environments.

It should be noted that deep learning based approaches claim the state-of-the-art results for unconstrained face recognition [30]–[32]. Such approaches have significantly improved face recognition in the wild [48]. Overall, in the presence of traditional covariates the recognition performance has matured, however, face recognition in fully uncontrolled environments and with emerging covariates warrants a significant research attention [49], [50].

1.1.2 Scale of Data: Incremental and Large Scale Training

As adoption of face recognition systems in real world applications increases, so does the operating scale of such systems. The practical drivers for the challenge of scale include national identity projects, advances in surveillance systems, detailed biometric recording of suspects/offenders, and biometrics based transactional authentications. Further, need of incorporating various representations of face (multi-spectrum, multi-resolution, hand-drawn and composite sketches) also eventually contribute to broadening the operational scale of face recognition systems. On one hand, large scale data opens possibilities to learn better models, whereas on other had, it adversely affects the training and query processing time.

It is well observed that more accurate models may be obtained by leveraging large training sets. Unfortunately, most of the efficient learning algorithms, such as Support Vector Machine (SVM), have super-linear time and space complexity with respect to training set sizes and feature dimensionality. Due to this property, most of the learning algorithms scale poorly with massive training sets. In terms of run-time query processing time, identification (1:N matching) and de-duplication tasks have time complexity directly proportional to the enrollment set size. For example, in Aadhaar project[6] the de-duplication needs to be performed for the population size of whole nation (approximately 1.2 billion people for India). Similar challenge of 1:N matching is also posed in surveillance scenarios. Considering these challenges, the academic community has kept on creat-

[5] https://www.nist.gov/programs-projects/face-recognition-prize-challenge, last retrieved: 9 July 2017

[6] https://uidai.gov.in/beta/

Table 1.2: Summary of various public face datasets. Notice the increase in the size (number of subjects and samples) of the datasets in recent years.

Dataset	Year	Number of Subjects	Number of Samples
AT&T [51]	1994	40	400
Color FERET [53][54]	2001	994	11,338
PIE [55]	2002	68	41,368
ND-Collection B [56]	2003	487	33,287
FRGC 2.0 [57]	2005	568	44,278
MORPH-II [58]	2006	13,673	55,608
LFW [59]	2007	5,749	13,233
PubFig [60]	2009	200	58,797
CMU-MultiPIE [61]	2010	337	755,370
YTF [62]	2011	1,595	3,425 videos
WDRef [28]	2012	2,995	99,773
MSRA-CFW [63]	2012	1,583	202,792
PaSC [64], [65]	2013	293	9,376 images, 2,802 videos
FaceScrub [66]	2014	530	107,818
CASIA-WebFace [67]	2014	10,575	494,414
Celeb-Faces+ [30]	2014	10,177	202,599
MegaFace [52]	2015	690,572	1,027,060
VGG-Face Dataset [68]	2015	2,622	982,803
IJB-A [46]	2015	500	5,712 images, 2,085 videos
IJB-B [47]	2017	1,845	21,798 images, 7,011 videos
CrowdFaceDB [69]	2017	257	384 videos

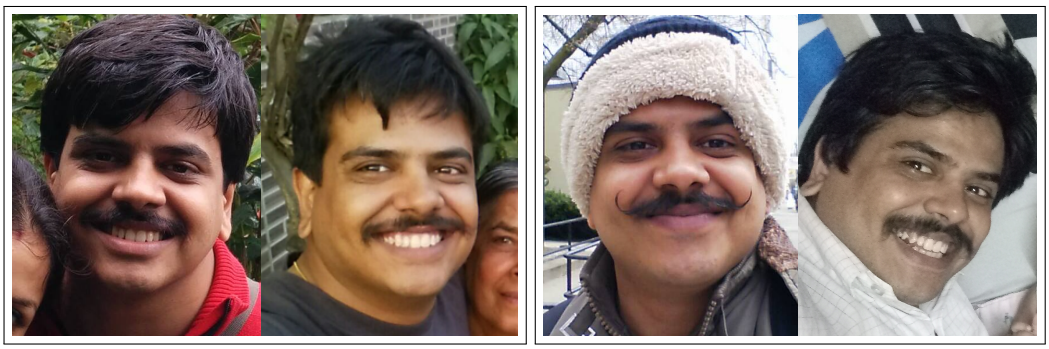
ing larger datasets over the years. Some of the important face datasets are listed in Table 1.2 along with their sizes. Notice that in 20 years, research community has moved forward from dataset of 40 subjects [51] to 0.7 million subjects [52]. The increase in dataset sizes is indicative of the need for practical scalable face recognition systems.

Additionally, in many scenarios the data may not even be available in one batch, e.g. when a repeat offender is caught/booked after a long time, the identification system has to be updated with the new *incremental* face samples acquired from him/her. To illustrate it further, Figure 1-6 shows images of the same subject acquired between years 2010 and 2017. Assume that the suspect is first caught in 2010, he is enrolled into the recognition system with two images, and is later released. In 2011, his face samples are acquired again as part of a routine observation and reporting. In 2012, a face is captured in a CCTV feed and we need to establish the person's identity. Notice that the suspect's face appears more similar to the sample acquired in 2010 than 2011. Therefore, the

(a) 2010 (b) 2011

(c) 2012 (d) 2013

(e) 2014 (f) 2015



(g) 2016 (h) 2017

Figure 1-6: Illustrating the need to update the recognition engine with incrementally available face samples as the appearance may change. The image shows face images of a subject acquired over several years. Updating such (intra- and inter-) class characteristics in the classifier models is necessary.

chances of correctly identifying the suspect's face is likely to increase if the recognition system has been updated with the 2011 sample. The paradigm of learning in which the samples are not available in one batch and are required to be incorporated into the model in successive batches, is called as *incremental* learning.

Incremental learning can also be useful in dividing large scale learning problem into subproblems. In the sample images shown in Figure 1-6, the variations in person's face can be attributed to various factors including, aging, weight gain/loss, growth/removal of facial hair, and usage of facial accessories. However, with focus on scale and incremental learning, the goal is to develop scalable face recognition approaches independent of the covariates.

1.2 Role of Classifiers in Face Recognition Pipeline

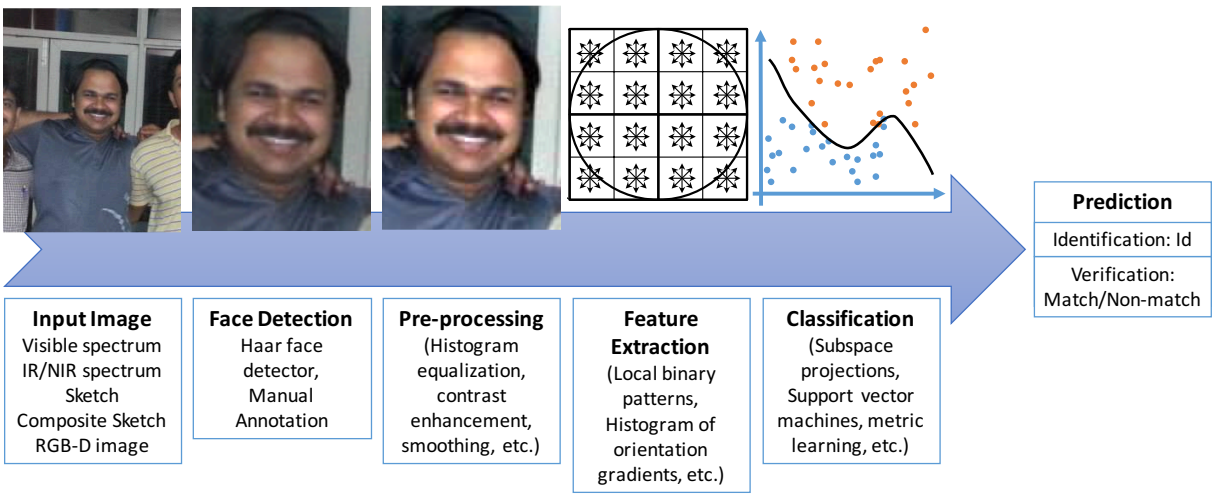


Figure 1-7: A typical face recognition pipeline.

As illustrated in Figure 1-7, a face recognition system, generally, contains face and landmark detection [70]–[72], pre-processing [73], feature extraction, and classification modules. Face region is first detected from the input image. The detected face is pre-processed to make it more suitable for further processing. Pre-processing stage typically involves geometric normalization to fit a detected face to a canonical face frame and image enhancement such as histogram equalization, masking, and smoothing. The pre-processed face image is provided as input to the feature

extraction module. Features are extracted to obtain holistic and/or local representation of face characteristics. In literature, various texture, gradient and learning-based features are widely explored in feature extraction module. The output of feature extraction module is, generally, a vector representation of the input face image. In the training phase, the classifier module utilizes these feature vectors corresponding to training set images to learn a classification model. In Figure 1-7, the functionality of classifier is illustrated as learning the classification decision boundary. During the testing phase, the feature vector representation of the query face image (or pair of face images) is fed into the classifier to obtain a predicted class label. For an identification system, the output is the predicted identity label. For a verification system the output is match or non-match.

The effectiveness of a face recognition system depends on how well the features encode the identity information and how well the classifier is trained. Usually, it is a trade-off between the complexity of feature extractor and the classifier. Utilizing features that are robust to the covariates can help in reducing the complexity of classifier design. Vice versa, a complex classifier many not heavily rely on invariant feature representation. Nonetheless, the overall best performances are generally reported with sophisticated features in conjunction with accurate classifiers. For example, a lot of deep learning related research efforts are focused on learning effective features/representations. However, these approaches tend to employ sophisticated classifiers along with deep features. Designing invariant features is heavily domain dependent, whereas designing variation-aware classifiers is largely domain independent as the later addresses the abstract/generic problem of underlying covariates. Thus, it allows to develop generic approaches which may be adapted for other classification tasks too.

In the chosen scope of face recognition in law enforcement, the unconstrained environment related challenges may be addressed at feature extraction and/or at classification modules. However, the challenges pertaining to the scale of data usually originate at the classifier level as the time and memory requirements are super-linear. Therefore, this dissertation addresses the challenges by improving existing classifiers and designing new ones for specific tasks.

1.3 Research Objectives and Contributions

The aim of this dissertation is to advance the face recognition research for law enforcement applications. As discussed earlier, the challenges in the given scope of problem are in two broader facets: uncontrolled environment and scale of data. The dissertation makes four major contributions with respect to uncontrolled environments and scale of data. Specifically, the research objectives and the contributions made towards them are:

1. **To develop algorithm for recognizing disguised faces aided by human evaluation study.**

   Traditionally face recognition research has seldom focused on mitigating the challenges of adversarial user behavior. This research focuses on one such adversarial behaviour of *disguise*. The aim is to understand and improve the performance of identifying disguised faces. We investigate human and machine performance for recognizing/verifying disguised faces. The performance is evaluated under factors of familiarity and match/mismatch with the ethnicity of observers. The findings of this study are used to develop an automated algorithm to verify the faces presented under disguise variations. We use localized feature descriptors which can identify disguised face patches and account for this information to achieve improved matching accuracy. In the proposed approach, the classification module involves disguised patch detection and feature comparison. The performance of the proposed algorithm is evaluated on the IIIT-Delhi Disguise database that contains images pertaining to 75 subjects with different kinds of disguise variations. The experiments suggest that the proposed algorithm can outperform a popular commercial system and equates to human performance for matching disguised face images.

2. **To develop algorithm for cross-view face recognition: cross-spectrum, cross-resolution, and photo-to-sketch**

   In law enforcement applications, it is important to mitigate the challenges posed by heterogeneity of face representations such as spectrum and resolution variations. The heterogeneity of spectrum and resolution are observed in surveillance application. Similarly, suspect identification scenario involves heterogeneity due to the need of comparing sketches against photographs. The objective is to improve the face recognition under influence of

such heterogeneity. In this research, we present two heterogeneity-aware subspace techniques, Heterogeneous Discriminant Analysis (HDA) and its kernel version (KHDA) that encode heterogeneity in the objective function and yield a suitable projection space for improved performance. We next propose the face recognition framework that uses existing facial features along with HDA/KHDA for matching. The effectiveness of HDA and KHDA is demonstrated using handcrafted and learned representations, on three challenging heterogeneous cross-view face recognition scenarios: (i) visible to near-infrared matching, (ii) cross-resolution matching, and (iii) digital photo to composite sketch matching. Comparison with state-of-the-art heterogeneous matching algorithms shows that HDA and KHDA based matching yield state-of-the-art results on all three case studies.

3. **To develop algorithm for efficient incremental updating of subspace learning based face recognition model.**

   It is possible that not all the training face images are available in one single batch for learning the classification model, or it may not be possible to learn from the entire large-scale training data due to hardware and/or algorithmic limitations. This presents the challenge of learning classifiers or matchers from batches of training data available incrementally. The challenge is particularly severe if the recognition pipeline involves subspace based approaches. Thus, the objective is to learn accurate models from such incremental data in a time efficient manner.

   The research focuses on developing a subspace approach that can incrementally update the model. A computationally effective incremental update procedure is devised that can leverage unlabeled data. In the proposed approach, total scatter matrix is estimated offline using unlabeled data whereas only the between-class scatter matrix is updated using incremental data. The effectiveness of the proposed algorithm, termed as *Incremental Semi-Supervised Discriminant Analysis* (ISSDA), is evaluated for face recognition application. The performance is evaluated, using CMU-PIE, CMU-MultiPIE, and NIR-VIS-2.0 face datasets, by comparing the accuracy, time and consistency of the proposed incremental algorithm with the corresponding batch learning models. Evaluations to understand the effects of the manifold regularizer and unlabeled data size are also performed. Further, the effect of updating the model with incremental batch consisting of samples of new classes is also studied.

4. **To develop computationally efficient algorithm to learn support vector machine for large-scale face recognition**

   SVM is considered amongst one of the best performing classifiers for a variety of tasks. However, its time complexity hinders learning from large scale training data. This limitation is also an impediment for employing and learning SVM based face recognition systems in conjunction with large scale training. Thus the objective is to improve SVM learning procedure/solver to enable faster and efficient, yet accurate learning.

   We propose an approach for learning kernel Support Vector Machines from large-scale data with improved computation time. The proposed approach, termed as *Subclass Reduced Set SVM* (SRS-SVM), utilizes the subclass structure of training data to effectively estimate the candidate support vector set. As the candidate support vector set cardinality is only a fraction of the training set cardinality, learning SVM from the former requires less time; without significantly changing the decision boundary. Further, we also propose the hierarchical extension of SRS-SVM. The effectiveness of the proposed approach is evaluated on five synthetic and six real world datasets. The qualitative analysis of the decision boundary, as well as extensive quantitative analysis of computation time, classification accuracy, precision-recall of the candidate set estimation, and effect of parameters is presented. On various datasets the SRS-SVM yields similar classification accuracy while requiring few folds reduced computation time as compared to traditional solver (LibSVM) and state-of-the-art approaches such as divide-and-conquer SVM, FastFood, and LLSVM.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Recognizing Disguised Faces

Figure 2-1: Illustrating the effect of disguise accessories on inter-class and intra-class variations.

Disguise is an interesting and a challenging covariate of face recognition. It involves both intentional and unintentional changes on a face through which one can either obfuscate his/her identity and/or impersonate someone else's identity. In either case, facial disguise falls under the broader category of *biometric obfuscation* [74]. Figure 2-1 shows an example of *face obfuscation* where the appearance of a subject can vary by using different disguise accessories. (Note that the images in Figure 2-1 may be affected by covariates other than disguise, e.g. aging; however, in this work we are concentrating on disguise only). As shown in Figure 2-1, disguise increases the *intra-class* variation (when it is used to hide one's identity) and reduces the *inter-class* variation (when it is used to impersonate someone else). Even though the problem of face recognition

under disguise is prevalent in real world applications, it has not been studied extensively. To make automatic face recognition more usable and secure, it is necessary to address the problem of (at least unintentional) disguise.

In recent years, recognition of disguised faces by humans has been an interesting area of research for cognitive scientists. Righi *et al.* [75] studied the effect of adding or removing the disguise accessories such as wigs and eyeglasses. They also evaluated the *switch*/*no switch* scenario where the accessories present during training phase were removed (switch) or kept unaltered (no switch). The study revealed that increasing the alterations to facial attributes of the probe image decreased the recognition performance. Further, the change in the rather stable facial features such as eyes had comparatively higher impact in decreasing the performance. A more detailed analysis regarding the effect of disguise on eye region was presented in [76], [77]. Sinha *et al.* [76] studied the importance of eye brows stating "Of the different facial features, eyebrows are among the most important for recognition". Douma *et al.* [77] found that removing glasses during testing had more damaging effect than adding; this is also called as the Clark-Kent effect [78]. A significant recognition performance difference was observed with variation in degree of familiarity, i.e. familiarizing the participant nine times did show significant performance difference than familiarizing three times. At a level of abstraction, Sinha *et al.* [76] and Douma *et al.* [77] provided insights about the effect of disguise on stable features. Complimentarily, the effect of hair – rather unstable features – was studied by Toseeb *et al.* [79]. The authors observed no significant performance difference when the participants were shown faces with and without hair. The phenomenon was attributed to the *internal face features*, which remained constant in both the scenarios. Similarly, the effect of internal features was also studied in [80], [81]. Overall, it appears that the effect of disguise on stable facial parts has more impact than on the unstable facial parts. However, to the best of our knowledge, a comprehensive research on the effect of disguising individual facial parts and their combinations is not performed.

Since disguise can be viewed as alteration to visual face information, the research related to recognition of altered/degraded facial images can potentially provide some insights. In presence of image degradation by blurring, Sinha *et al.* [76] have shown that familiarity of the stimuli subjects is advantageous for face recognition. Complimentarily, Hancock *et al.* [82] reported that unfamiliar faces are difficult to recognize in a low-quality surveillance video. Combining their

results [76], [82] point to a possibility that the representation of familiar faces might be more robust to certain image degradations than that of unfamiliar faces. Therefore, understanding the effect of familiarity on disguised face recognition can potentially provide insights into the robust facial representation and recognition by humans. It has been also observed in literature that face recognition by humans is subjective to familiarity [83] and race [84].

A brief overview of literature related to automated face recognition under disguise variations is presented in Table 2.1. Note that most of the research has been performed using the AR [40] and Yale [90] face databases which contain very limited disguise (sunglasses and scarves only). However, to be confident about the performance of automated approaches, it is required that evaluation is performed on a dataset with more exhaustive disguise variations. Regarding the effect of ethnicity, Phillips *et al.* [100] evaluated the performance of algorithms on east Asian and Caucasian faces. The study showed that the fusion of the algorithms developed in east Asia performed better on east Asian faces than on Caucasian faces. Similarly, fusion of the algorithm developed in West countries performed better on Caucasian faces than east Asian faces.

In the last decade, some studies compared the performance of automated face recognition algorithms and humans. O'Toole *et al.* [101] compared human performances with academic and commercial systems. They observed that on the *easy* pairs, all the automated algorithms, except one, exhibited better performance than humans; while for the *difficult* pairs, some algorithms outperformed humans. This study focused on understanding the effects of the illumination variation and, interestingly, the image pairs that were *difficult* for PCA based algorithms were also found to be difficult for humans. Moreover, the evidences of algorithms surpassing humans for face verification task were also observed. Similar comparison was presented in [102] for face recognition under uncontrolled illumination, indoor and outdoor settings, and day-to-day appearance variation. In [102], algorithms were shown to have superior performance than humans for *good* and *moderate* image pairs, whereas humans and algorithms were comparable for the *poor* accuracy group. These good, moderate, and poor accuracy groups were created based on scores given by algorithms. Though not for face recognition, but for face detection, Marius't [103] reported the *similar-error* phenomena by humans and automated algorithm (AdaBoost cascade classifier [70]). Further, O'Toole *et al.* [104] fused the humans and algorithms for face verification task using partial least square regression. The fusion resulted in significant performance improvement. To the

Table 2.1: **Literature review:** Existing algorithms for addressing disguise variations. AR database [40] contains 3200+ images pertaining to 126 subjects with two kinds of disguises (sunglasses and scarves). The National Geographic (NG) dataset contains 46 images of 1 individual, with various accessories such as hat, glasses, sunglasses, and facial hair. *Private dataset of 150 images pertaining to 15 individuals which contains similar real and synthetic disguise variations as in NG dataset. $^{+}$Synthetic disguise dataset of 4000 images pertaining to 100 individuals. $^{\times}$Private datasets are collected by researches in real world scenarios from ATM (automatic teller machine) kiosk.

Authors	Algorithm	Disguise detection	Disguise / occlusion detected as	Face recognition	Spectrum	Database
Ramanathan *et al.* [85]	PCA	Yes	Left/right half face	Yes	Visible	National Geographic, AR
Singh *et al.* [86]	2D-log polar Gabor	No	-	Yes	Visible	AR, Private*, Synthetic Disguise^{+}
Marsico *et al.* [87]	Partitioned iterated function system	No	-	Yes	Visible	AR
Shreve *et al.* [88]	Optical flow	No	Facial Strain Map	Visible	Video dataset	
Martinez [89]	Probabilistic matching	No	-	Yes	Visible	AR
Wright *et al.* [27]	SRC	No	-	Yes	Visible	AR, Yale B [90]
Kim *et al.* [91]	ICA	No	-	Yes	Visible	AR, FERET
Yang and Zhang [92]	Gabor SRC	No	-	Yes	Visible	AR, Yale B
Pavlidis and Symosek [93]	-	Yes	Not explicitly	No	Near-IR	-
Yoon and Kee [94]	PCA + SVM	Yes	Upper/lower half	No	Visible	AR, Private$^{\times}$
Kim *et al.* [95]	PCA + SVM	Yes	Upper/lower half	No	Visible	AR, Private$^{\times}$
Choi and Kim [96]	AdaBoost + MCT-based features	Yes	Left-right eye, mouth	No	Visible	AR
Min *et al.* [97]	Gabor + PCA + SVM, LBP	Yes (Gabor + PCA + SVM)	Upper/lower half	Yes (LBP)	Visible	AR
Dhamecha *et al.* [41]	ITE, LBP	Yes (ITE)	Individual patches	Yes (LBP)	Visible and Thermal	I^2BVSD
Yang *et al.* [98]	Nuclear Norm based Matrix Regression	No	-	Yes	Visible	AR

best of our knowledge, neither 1) a study focusing on covariate of disguise has been carried out, nor 2) any attempt to enhance machine performance by encoding human strategy for recognizing disguised faces has been made.

In this research we evaluate the effect of familiarity and ethnicity on disguised face recognition, and attempt to encode learnings from human evaluations into an automated algorithm. Since humans are considerably efficient at face recognition [101], comparison of humans and automated algorithms is also performed. The main contributions from this research can be summarized as follows:

- evaluating human face recognition performance under face disguise along with familiarity and ethnicity/race effect;
- determining the effect of individual facial parts on the overall human face recognition performance;
- proposing an automated face recognition algorithm based on the learnings from human evaluation and comparing the performance with Sparse Representation Classifier (SRC) [27] and a commercial off-the-shelf (COTS) system; and
- comparison of human performance with automated algorithms (including the proposed algorithm) for addressing disguise variations.

2.1 Material and Methods

2.1.1 Ethics

To undertake this research the first step was to create a database. At the time of database creation all the 75 subjects in the database were of age 18+ years. The subjects were provided with accessories, and were asked to use the accessories at their will in order to get disguised. All the subjects provided written informed consent for using their face images for research purpose. The consent, for sharing their face images with research community and publish their face images in research papers, was also taken from the subjects. Images pertaining to only those subjects who gave their consent for sharing their face images, will be made available to the research community.

Face Recognition under Disguise: Human Evaluation

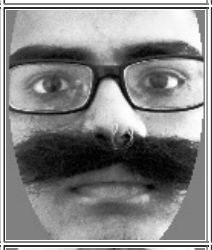
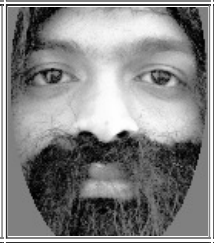
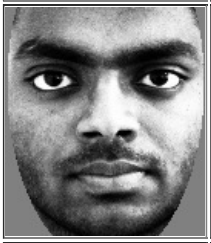
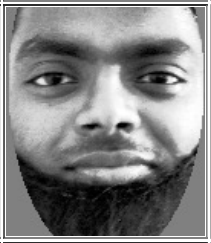
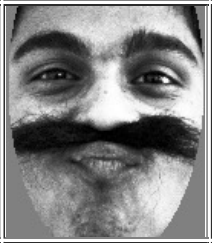
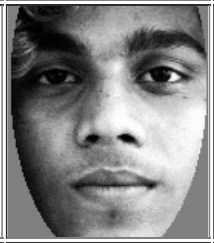
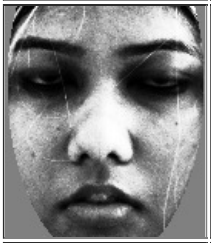
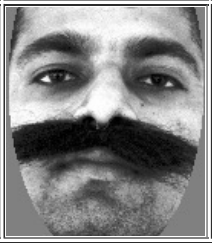
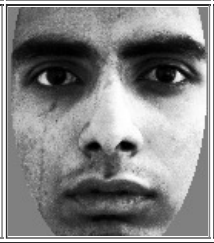
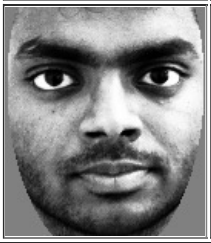
Questionnaire Number=1

Instructions:

You have to answer questions 1 to 8 to determine whether the image pairs are of the same person or not. Providing age and gender information is optional.

Gender:_____[optional]

Age:_____[optional]

No.	Image 1	Image 2	Same Person? Yes(✓)/No(✗)	No.	Image 1	Image 2	Same Person? Yes(✓)/No(✗)
1			☐	2			☐
3			☐	4			☐
5			☐	6			☐
7			☐	8			☐

Declaration:

By filling this survey form, you agree that your responses can be used for research purposes.

Figure 2-2: Sample questionnaire.

In order to analyze human capability of recognizing disguised faces, we collected the responses from various participants. All the responses collected from survey participants are anonymous and are used only for research purposes. Their willingness to participate in the survey was also asked. A sample survey collection form is shown in Figure 2-2. The database collection and survey response collection procedures for this study were approved by the IIIT-Delhi Ethics Board.

2.1.2 Disguise Face Database

Figure 2-3: Sample images from the ID V1 database.

The databases generally used for disguise related research (AR [40] and Yale [90] face databases) contain very limited disguise variations, such as scarves and/or sun-glasses. Therefore, to evaluate the effectiveness of automated algorithms and to evaluate human performance, we have collected the IIIT-Delhi Disguise Version 1 face database (ID V1) of disguised/obfuscated face images. The ID V1 database contains 681 visible spectrum images of 75 participants (all above the age of 18 years) with disguise variations. The number of images per person varies from 6 to 10. For every subject, there is at least one frontal neutral. Here, face image without any disguise is referred as *neutral* face image. face image and at least five frontal disguised face images. All the face images are captured under (almost) constant illumination with neutral expression and frontal pose. The disguise variations included in the database are categorized into the following categories.

- **Without disguise**: neutral image,
- **Variations in hair style**: different styles and colors of wigs,
- **Variations due to beard and mustache**: different styles of beards and mustaches,
- **Variations due to glasses**: sunglasses and spectacles,
- **Variations due to cap and hat**: different kinds of caps, turbans, veil (also known as hijab which covers hair), and bandanas,
- **Variation due to mask**: disposable doctor's mask, and
- **Multiple variations**: a combination of multiple disguise accessories.

Figure 2-3 shows sample images from the database. The disguises are chosen in such a way that they result in more realistic appearances and (almost) every part of the face is hidden at least once. The subjects are asked to disguise themselves using the given accessories. This allows different subjects to have different types of disguises thus providing more variations across individuals in the database. The database is publicly available for research purpose [1]. The images from the dataset are preprocessed in the same way as in [41] i.e. preprocessing is done using the CSU Face Identification Evaluation System [105] to obtain normalized images.

2.1.3 Participants for Human Evaluation

Since this study examines the effect of ethnicity and familiarity factors on face recognition with disguise variations, the participants were divided into the following four sets.

Set 1: familiar to the subjects in Stimuli and of the same ethnicity as subjects (Set FS-I),

Set 2: familiar to the subjects in Stimuli and of the same ethnicity as subjects (Set FS-II) (redundant set of Set 1),

Set 3: unfamiliar to the subjects in Stimuli and of the same ethnicity as subjects (Set US), and

[1] IIIT-Delhi Disguise Version 1 face database is available at https://research.iiitd.edu.in/groups/iab/resources.html

Set 4: unfamiliar to the subjects in Stimuli and of different ethnicity than subjects (Set UD).

Note that, one more combination, i.e. familiar to the subjects in Stimuli and of different ethnicity, is possible. However, due to the lack of participants satisfying this criteria, we have not been able to show study related to such a set.

2.1.4 Stimuli, Design and Procedure

Each of the four sets consisted of 100 unique participants and the stimuli consisted of subjects of ID V1 dataset collected at IIIT-Delhi. Since the participants in Sets FS-I & FS-II and stimuli belonged to the same department in IIIT-Delhi, it ensured familiarity and same ethnicity factors. Set FS-I and Set FS-II were redundant in nature, as they were similar in terms of familiarity and ethnicity. However, having access to two groups with participants of same variable provided scope for more analysis in terms of the consistency of outcomes. To ensure the unfamiliarity factor in Set US, it consisted of participants from another city of a different state of India. As the two cities are far apart and no logical connection among subjects and participants was known, it was safely assumed that the participants in Set US were unfamiliar to the stimuli subjects. Since the participants in Set FS and Set US were from India, they were of the same ethnicity as the stimuli. Set UD consisted of participants of Chinese ethnicity, thus ensuring unfamiliarity and different ethnicity than that of stimuli. Table 2.2 summarizes the details regarding the number of participants and gender distribution in each set.

Each participant was given a questionnaire containing eight face image pairs. The participants were supposed to mark them as “same person" or “not same person". Optionally, the participants were also requested to write their age and gender. Each participant in a set was given a unique questionnaire. However, there were overlapping questions among different questionnaires. Therefore, 100 questionnaires were designed by randomly choosing genuine (same person) and impostor (different person) image pairs with equal priors. The pairs were drawn from a split that contained neutral and disguised face images pertaining to 40 subjects. The pairs for each questionnaire were selected with substitution, therefore an image pair could appear in multiple questionnaires; however it was made sure that no image pair was repeated in the same questionnaire. Thus, across 100 questionnaires, 436 unique image pairs were used. Figure 2-4 shows the distribution of genuine

Table 2.2: **Age and gender distribution of participants in the four sets.** The results reported are mean values with standard deviation.

Set	Overall		Male		Female		Gender Not Specified
	No.	Age $\mu \pm \sigma$	No.	Age $\mu \pm \sigma$	No.	Age $\mu \pm \sigma$	No.
Familiar, Same Ethnicity-1 (FS-I)	100	18.5 ± 0.8	68	18.5 ± 0.6	30	18.5 ± 0.6	2
Familiar, Same Ethnicity-1 (FS-II)	100	20.5 ± 3.5	58	20.7 ± 3.8	38	20.2 ± 3.8	4
Unfamiliar, Same Ethnicity (US)	100	19.5 ± 2.5	64	19.5 ± 2.5	33	19.5 ± 2.5	3
Unfamiliar, Different Ethnicity (UD)	100	23.6 ± 3.8	55	24.6 ± 5.6	44	22.4 ± 5.6	1

and impostor pairs in questionnaires. Note that the majority of questionnaires had an even mixture of genuine and impostor image pairs. Further, the face images were converted to gray scale and elliptical mask was applied to face images to make sure that no features other than facial cues could be used for recognition. All the face images were resized to 130×150 pixels which translated to 2.8cm$\times$3.2cm on a printed document of A4 size. One such example questionnaire is shown in Figure 2-2. The exact same set of 100 questionnaires was used for collecting responses from the participants of Set FS-I, Set FS-II, Set US, and Set UD.

One of the objectives of this research is to compare human evaluation with automated algorithms. Automated algorithms are generally evaluated in either face matching/verification or face identification scenarios. In face matching or verification scenario, an image pair is classified as match or non-match, whereas in face identification scenario a query image is compared with gallery/enrolled face images to predict the identity. For comparing the human and machine performance, it is essential that the comparison metric is same for both. Simulating identification scenario for human evaluation involves two challenges:

- First, the gallery images are to be shown to the subjects for *enrolling* them in their memory. However, this process becomes challenging with increasing number of gallery images.

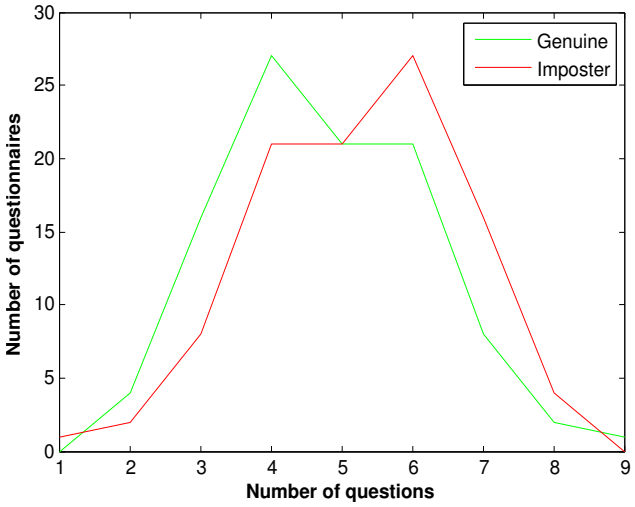


Figure 2-4: Distribution of genuine and impostor pairs in questionnaires.

- Identification performance of an automatic algorithm is measured in terms of cumulative match characteristics (CMC) curve, which requires to get ranked list of gallery images in sorted order of matching with the query image. Therefore, if human performance is to be compared with algorithms in identification scenario, the ranking is required to be generated by humans too. This is practically possible if number of gallery images is small. However, it is rather difficult, from experimental design and participants perspective, when the number of gallery images is large.

Further, existing research in human versus algorithm comparisons focuses on verification scenario [101], [102]; therefore this paper also focuses on the same. Apart from comparing, we also aim at incorporating the understandings from human cognition into an automated algorithm.

A mixed-subjects design was employed in which the *between-subjects* variables were familiarity (familiar or unfamiliar), ethnicity (same as stimuli or different from stimuli), and gender (male or female). The participants took part in only one of the four sets/Familiarity-Ethnicity combinations (Set FS-I, Set FS-II, Set US, and Set UD). The combination of Familiar-Different Ethnicity could not be evaluated as it is challenging to find such participants. The *within-subjects* variable was the amount of disguise on stimuli face images. The participants in each of the sets followed the same procedure, i.e. they were given a questionnaire containing eight face image pairs and they marked each pair as “same person" or “not same person".

The evaluations are performed in terms of the False Accept Rate ($\mathrm{FAR} = 100 \times \frac{\mathrm{FA}}{(\mathrm{FA+GR})}$), Genuine Accept Rate ($\mathrm{GAR} = 100 \times \frac{\mathrm{GA}}{(\mathrm{GA+FR})}$), and Accuracy ($\mathrm{Acc} = 100 \times \frac{(\mathrm{GA+GR})}{(\mathrm{GA+FA+GR+FR})}$), where $\mathrm{GA}, \mathrm{FA}, \mathrm{GR}$ and FR represent the number of genuinely accepted, falsely accepted, genuinely rejected, and falsely rejected pairs respectively. False accept means that a non-match pair is classified as a match pair and genuine accept means that a match pair is correctly classified. A face recognition is expected to achieve high GAR at low FAR.

The results of F-test with ν_1 and ν_2 degrees of freedom are denoted as $F(\nu_1, \nu_2)$, similarly, the t-test with ν degrees of freedom is denoted as $t(\nu)$. All the test results are reported with the corresponding p-value. $p < 0.05$, $p < 0.01$, and $p < 0.001$ indicate moderately, strongly, and very strongly significant evidences respectively. .

Table 2.3: **Summary of human performance.** It is reported in terms of mean FAR, GAR and accuracy in each of the four sets.

Set	FAR $\%(\mu \pm \sigma)$	GAR $\%(\mu \pm \sigma)$	Accuracy $\%(\mu \pm \sigma)$
Set FS-I	19.62 ± 4.54	74.47 ± 4.80	75.87 ± 3.93
Set FS-II	17.79 ± 4.45	69.27 ± 5.15	75.12 ± 4.04
Set US	18.85 ± 4.34	57.88 ± 5.16	69.50 ± 4.10
Set UD	24.20 ± 4.90	57.47 ± 5.26	66.00 ± 4.17

2.1.5 Observations from Human Evaluation

The responses collected from participants of all the sets (Set FS, Set US, and Set UD) are used to compute the false accept rate, genuine accept rate, and accuracy. The major reason for evaluating the FAR and GAR along with accuracy is that accuracy does not provide information about GAR and FAR individually. Therefore, evaluating the performance in terms of GAR and FAR separately may help in understanding the efficiency of matching genuine and impostor pairs individually. The mean and standard deviations are reported in Table 2.3.

Statistical tests are performed to further analyze these results. Three One-Way ANOVAs (Analysis of variance) are conducted to evaluate the statistical significance of FAR, GAR, and Accuracy. The results of these tests are as follows.

1. FAR ($\mathrm{F}(3, 396) = 1.82$, $p = 0.14$),

2. GAR ($\mathrm{F}(3, 396) = 10.54$, $p < 0.0001$), and

3. Accuracy ($\mathrm{F}(3, 396) = 8.08$, $p < 0.0001$).

This analysis of p-values shows that there is a significant difference in terms of GAR and accuracy with the corresponding $p < 0.0001$ for both the statistics. However, there is no significant difference for FAR, since $p = 0.14$. Post-hoc analysis is carried out using paired t-test to understand the 1) effect of familiarity, 2) effect of ethnicity, 3) effect of gender, 4) consistency between Set FS-I and Set FS-II, and 5) effect of specific disguise. The details of this analysis are provided below. The results and inferences of the statistical tests to understand the effect of familiarity, ethnicity, gender and consistency are summarized in Table 2.4.

Effect of Familiarity

To evaluate the effect of familiarity for each of the three statistics i.e. FAR, GAR, and Accuracy, two paired t-tests are performed: 1) between Set FS-I and Set US and 2) between Set FS-II and Set US. In both cases, significant accuracy improvement is observed when the participants are familiar to the stimuli. The p-values for accuracy are reported as follows.

- Set FS-I and Set US: t(99) = 2.99, $p = 0.0035$
- Set FS-II and Set US: t(99) = 2.80, $p = 0.0061$

However, no significant difference is observed for FAR.

- Set FS-I and Set US: t(99) = 0.288, $p = 0.7829$
- Set FS-II and Set US: t(99) = -0.4060, $p = 0.6856$

Further, GAR is observed to be different for both the cases

- Set FS-I and Set US: t(99) = 4.86, $p < 0.0001$
- Set FS-II and Set US: t(99) = 3.14, $p = 0.0022$.

It is observed that the best performance is achieved when the participants are familiar with the stimuli face and are of the same ethnicity. Interestingly, Sets FS-I & FS-II have the same FAR as

Table 2.4: p**-values of statistical tests to understand the effect of each factor.** ✓represents that the corresponding statistical test show significant difference between the compared sets and $\times$ represents insignificant difference.

Factor	Sets Compared	FAR	GAR	ACC	Inference
Familiarity	FS-I & US	0.7829 ($\times$)	<0.0001 (✓)	0.0035 (✓)	Unfamiliarity degrades GAR but not FAR
	FS-II & US	0.6856 ($\times$)	0.0022 (✓)	0.0061 (✓)	
Ethnicity	US & UD	0.0715 ($\times$)	0.9103 ($\times$)	0.0789 ($\times$)	No additional degradation
Consistency	FS-I & FS-II	0.6878 ($\times$)	0.1025 ($\times$)	0.7199 ($\times$)	Both sets are consistent
Gender	FS-I (M) & FS-I (F)	0.1573 ($\times$)	0.2420 ($\times$)	0.0171 (✓)	Female are better in Sets FS-I and FS-II. For other sets no significant difference is observed
	FS-II (M) & FS-II (F)	0.4529 ($\times$)	0.6801 ($\times$)	<0.0001 (✓)	
	US (M) & US (F)	0.3776 ($\times$)	0.3785 ($\times$)	0.9535 ($\times$)	
	UD (M) & UD (F)	0.3535 ($\times$)	0.2737 ($\times$)	0.1524 ($\times$)	

Set US, but Set US has significantly lower GAR. This means that when participants are unfamiliar to stimuli, they tend to reject more genuine matches. From the observation regarding similar FAR in Set FS-I, FS-II, and US, one can claim that: if a pair has images of different individuals, an unfamiliar participant will classify it as “same person" with equal likelihood as a familiar participant. Moreover, the finding that “familiar faces are easier to match even if they are disguised" is equivalent to the similar finding for non-disguised faces [83]. Although, Douma *et al.* [77] did not find the effect of familiarity significant in recognizing disguised faces, note that our experimental procedure is different from their's. In [77], the participants were to *identify* the stimuli faces, whereas in this study the participants were to classify the stimuli image pairs as “same person" or “different persons". The former involves the face identification scenario, where the performance is primarily a function of memory and internal representation of faces which is enhanced if the person is familiar. However, that is not the case with our study which involves face verification scenario as it enables us to compare human performance with algorithm. To summarize, *familiarity is an advantageous factor and unfamiliarity significantly degrades genuine accepts but not the false accepts.*

Effect of Ethnicity

To understand the effect of ethnicity, paired t-tests are performed between Set US (unfamiliar, same ethnicity) and Set UD (unfamiliar, different ethnicity). The participants in both these sets are unfamiliar to the stimuli subjects; Set US has the participants which are of same ethnicity as stimuli, whereas Set UD participants are of different ethnicity than stimuli. Among the unfamiliar participants, the one with different ethnicity does not result in significantly different accuracy (t(99) = -1.7757, $p = 0.0789$). From further analysis in terms of FAR and GAR it is found that neither FAR (t(99) = 1.82, $p = 0.0715$) nor GAR (t(99) = -0.1129, $p = 0.9103$) is significantly differing. This suggests that in the presence of disguise, different-ethnicity factors do not add to the reduction in performance due to unfamiliarity factor. Therefore, the other-race effect [84] does not significantly further deteriorate the performance of recognizing disguised faces if the participants are unfamiliar to stimulus. However, if the participant is of the same ethnicity as the stimulus, familiarity is an added advantage.

Effect of Gender

No specific effect of gender is observed, except on the accuracy of Set FS-I (t(96) = -2.427, $p = 0.0171$) and Set FS-II (t(94) = -15.56, $p < 0.0001$) where female participants exhibit significantly better performance than male participants. However, even for these two sets no significant difference in FAR or GAR is observed. Similar observation regarding female superiority for face recognition has been studied in literature [106]. However, for disguised face recognition, this effect is observed only when the participants are familiar to stimuli faces and it disappears with absence of familiarity and/or difference in ethnicity.

Consistency between Set FS-I and Set FS-II

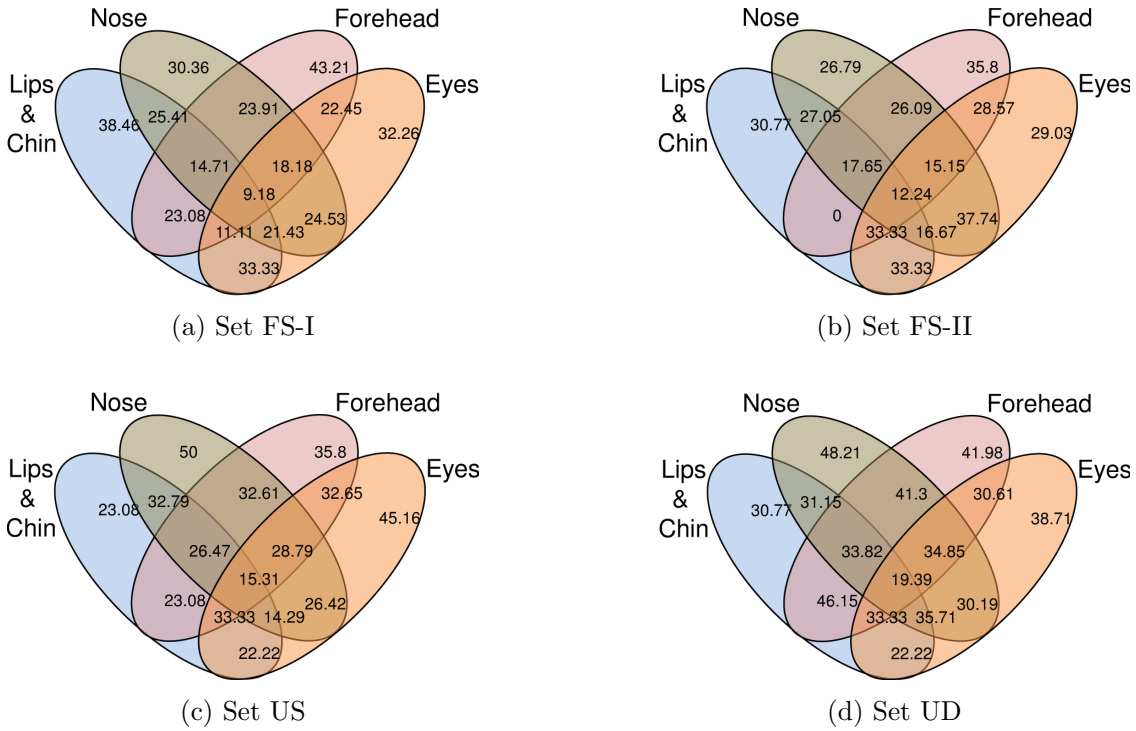
As we have access to two sets with the same familiarly and same ethnicity settings, it enables us to perform a consistency check, i.e. to evaluate similarity between the results of two sets with same design variables. We performed paired t-tests between Set FS-I & Set FS-II to analyze if there is any performance difference. Without much surprise, there is no significant difference in FAR (t(99) = 0.6878, $p = 0.4932$), GAR (t(99) = 1.6481, $p = 0.1025$), and accuracy (t(99) = 0.3596, $p = 0.7199$). For comparison, the response of both the sets are illustrated in the form of a confusion matrix in Table 2.5. Thus, similar performance is observed in both the sets.

Table 2.5: **Confusion matrix for comparing the consistency of Set FS-I and Set FS-II.** ✓and × represent the genuine and impostor classified samples respectively. The numbers in every cell represent the co-occurrence of decisions (correct/incorrect). For example, ✓✓block shows that for 227 image pairs, participants in both Set FS-I and Set FS-II responded that they were genuine pairs.

Confusion Matrix		Set FS-I	
		✓	×
Set FS-II	✓	227	108
	×	130	335

Effect of Specific Disguises

In this analysis, we focus on enhancing the understanding regarding the effect of specific kinds of disguises on face recognition performance. Human performance decreases when faces are dis-



(a) Set FS-I (b) Set FS-II

(c) Set US (d) Set UD

Figure 2-5: **Effect of disguising individual facials parts and their combinations.** The numbers represent the percentage of the misclassified face image pairs belonging to the corresponding disguise combination. For example, there are 31 image pairs with disguise on eye strips only, out of which 10 are misclassified by the participants in Set FS-I (a). This leads to the aforementioned incorrect classification fraction of $\frac{10}{31} \times 100 = 32.26\%$.

guised [75]. However, the effect of various kinds of disguises and their combinations is not yet well explored. The presence of disguise on certain facial parts can corrupt or occlude the partial face information thus degrading the face recognition performance. We divide the face image into uniform 5×5 grids and label the first, second and third rows as forehead, eyes, and nose regions respectively. The remaining two rows taken together are labeled as lips and chin region. From manual annotation of every rectangular patch of the grid, we have information regarding which patch contains disguise. The disguised patches are referred to as non-biometric patches. A region is considered to contain disguise if more than half of the patches in that region are non-biometric patches. Since the face images are divided into four non-overlapping regions, there can be (2^4=)16 combinations of disguised regions. These combinations can be represented in the form of a 4 set venn diagram. Figure 2-5 represents such a venn diagram representing the percentage of incorrectly classified face image pairs belonging to each disguise combination. Figure 2-5(a), (b), (c),

and (d) represent venn diagrams pertaining to Set FS-I, Set FS-II, Set US, and Set UD respectively. Note that in the ideal case, all the numbers in the venn diagram would be zero, i.e. none of the face image pairs belonging to any of the disguise combination is incorrectly classified. The key observations are as follows.

- Intuitively, the accuracy of disguised face recognition should reduce with increase in the amount of disguise. However, consistently for all four sets, considerably high errors are reported even when only a single kind of disguise is present (see the *only nose, only eyes, only forehead*, and *only lips* in Figure 2-5). This may be due to the fact that when an image-pair contains only one kind of disguise, one or both the face images contain similar kind of disguise. Also from the database section it can be noted that the number of disguise accessories applicable to each facial part, such as eye-glasses and bandanas, are limited in number. Therefore, variations in accessories disguising each facial part are limited. As the disguise accessories are encoded as part of the overall presentation in human perception [75], use of 1) same kind of disguise accessories among different users and 2) different kinds of disguise accessories on the same user might be leading to higher error rates.

- In the other regions of the venn diagram i.e. with multiple disguises, images in the face image pairs can have disguise accessories affecting different facial feature(s), therefore the argument regarding the similar disguise accessories cannot be applied to them.

- Intersecting areas of venn diagrams corresponding to facial hairs and wigs i.e. forehead-nose and forehead-nose-lips-and-chin also yield considerably high error rates, implying that the co-occurrence of wig and mustache (and beard) makes it challenging to match two faces. Though, the negative impact of combination of disguises is less prominent than that of disguise in only one part, there is a steady trend of its increased impact with increase of challenging factors, i.e. Set FS $\rightarrow$ Set US $\rightarrow$ Set UD.

2.2 Anāvṛta: Proposed Face Recognition Approach

From the human evaluation study presented above, it is clear that use of disguise accessories degrades the recognition performance. This is majorly because disguise accessories get encoded as

a part in the overall presentation [75]. Moreover, use of disguise accessories can also reduce the uniqueness of subjects. From automated face recognition point of view, Pavlidis and Symosek [93] have suggested that detecting disguise is necessary to efficiently recognize disguised faces. Therefore, using learnings from the human analysis, we develop the following hypothesis for automated face recognition:

"The facial part or patches which are under the effect of disguise (or occluded in most of the cases), are the least useful for face recognition, and may also provide misleading details. It is this misrepresentation that a person uses to hide his/her own identity and/or to impersonate someone else."

Building upon this intuition, we propose a framework, termed as Anāvṛta, for recognizing faces with variations in disguise. As illustrated in Figure 2-6 there are two stages in the proposed framework:

1. **Patch Classification**: It comprises dividing face image into patches and classifying them into *biometric* or *non-biometric* classes.

2. **Patch based Face Recognition**: Biometric patches are matched using local binary pattern (LBP) based face recognition algorithm.

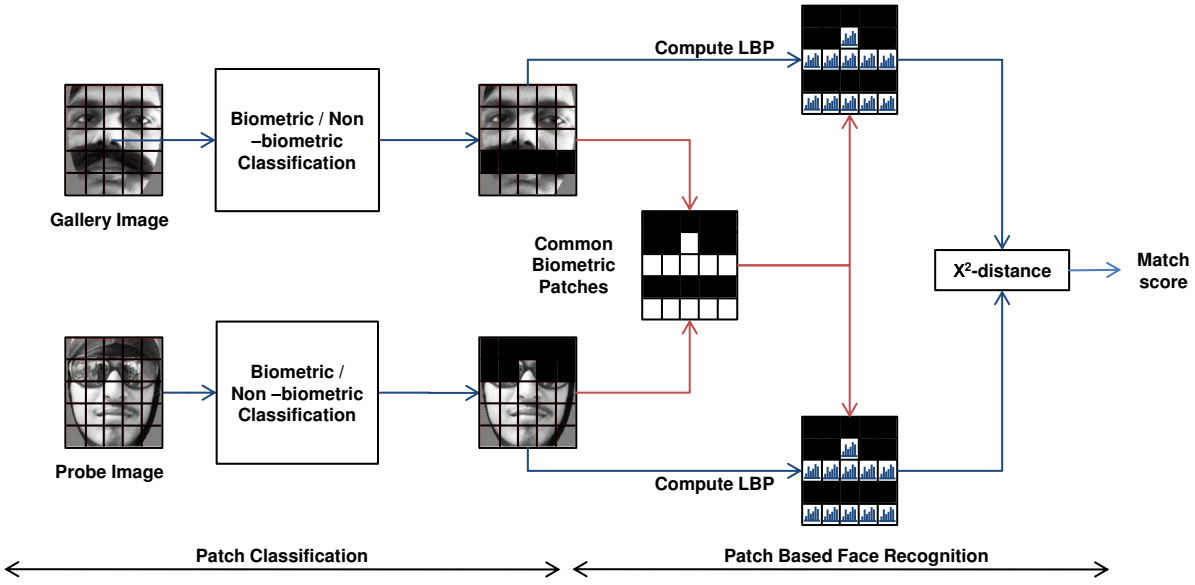


Figure 2-6: Illustrating the steps involved in the proposed face recognition framework.

2.2.1 Patch Classification

In human cognition research, Gosselin and Schyns [107] have proposed a technique to identify relevant facial regions for recognition which shows that certain facial parts are more important than others for recognition. In automated algorithm literature, several researchers have proposed patch or part-based face recognition [21], [108]–[110] and evaluated the performance of individual parts for face recognition. De Marsicso *et al.* [109], [110] proposed a solution based on local information where each facial part is used separately as input; the scores obtained by matching each part are fused to obtain final scores. Moreover, the mechanism for self-tuning the subsystems for matching individual parts was also proposed. To the best of our knowledge, [97], [111] are the only works in literature which use occlusion detection to enhance the recognition performance. In applications such as law-enforcement, analyzing the patches to determine whether they are genuine facial regions or accessories is very important. The proposed patch classification algorithm therefore aims to classify the patches into biometric and non-biometric classes.

- **Biometric patches** are those facial parts that are not disguised; and hence they are useful for recognition.

- **Non-biometric patches/artifacts** are facial parts that are disguised. These patches may reduce the performance and should be avoided as far as possible.

The patch classification algorithm has two steps: feature extraction and classification.

1. **ITE Feature Extraction:** It is our assertion that some of the non-biometric patches or occlusions, such as hair and artificial nose, can be distinguished using texture information, while some others, such as scarves and sunglasses, can be distinguished using their intensity values. Therefore, the proposed algorithm uses a concatenation of texture and intensity descriptors as input feature. As shown in Figure 2-6, the algorithm starts with tessellating the face image. Input face image I is first divided into non-overlapping rectangular patches I_{ij}, $1 \leq i \leq m, 1 \leq j \leq n$, where m and n are the number of horizontal and vertical patches respectively. The intensity and texture descriptors are computed for all the patches using the intensity histogram and Local Binary Patterns (LBP) algorithm [21] respectively. The

proposed descriptor is termed as the *Intensity and Texture Encoder* (ITE). For a patch ij of an image I, ITE is defined as

$$\mathbf{E}(I_{ij}) = [\text{intensityHist}(I_{ij}); \text{lbpHist}(I_{ij})] \tag{2.1}$$

where $\text{intensityHist}(\cdot)$ represents the histogram of an intensity image and $\text{lbpHist}(\cdot)$ represents the LBP histogram. We use basic LBP operator with 8 sampling points, that produces 256 dimensional feature vector for each patch. Intensity histogram consists of 256 bins, resulting in a feature vector of the same dimension.

2. **ITE Feature Classification:** The ITE features can, potentially, be classified using any of the generative or discriminative classification techniques. Our observation of biometric and non-biometric patches shows that the set of biometric patches is well defined and can be modeled efficiently. However, due to the variety of accessories that can be used for disguise, non-biometric patches have an exhaustive population set which is difficult to model using a limited training database. Therefore, in this research, we have used Support Vector Machine (SVM) [112], a discriminative classifier, for classifying biometric and non-biometric patches.

   An SVM model is learned from the ITE descriptors of all the patches from training images (which are annotated manually). This model is used to classify the patches from the testing data. For every patch, a score s is computed using SVM. A patch is classified as biometric, if the score is less than the threshold T, i.e. $s < T$; and if score is equal to or greater than the threshold, i.e. $s \geq T$, the patch is classified as non-biometric. Accordingly, a flag variable F_{ij} is generated, which represents whether the patch is classified as biometric or non-biometric. The flag value of every patch is then combined to generate the flag matrix, $\mathbf{F}_{m \times n} = [F_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n}$, using Eq. 2.2.

$$F(I)_{ij} = \begin{cases} 1 & \text{if } I_{ij} \text{ is classified as biometric} \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

   ITE features of images patches are classified using trained SVM model.

2.2.2 Patch based Face Recognition

Let I^p be the probe image which is to be matched with the gallery image I^g. The corresponding flag matrices $\mathbf{F}(I^p)$ and $\mathbf{F}(I^g)$ are generated using Eq. 2.2. Here, it is possible that for some gallery patch, I^g_{xy}, which is classified as biometric, the corresponding probe patch, I^p_{xy}, is classified as non-biometric. In other words, $F(I^g)_{xy} = 1$ and $F(I^p)_{xy} = 0$, or $F(I^g)_{xy} = 0$ and $F(I^p)_{xy} = 1$. This renders the particular patch of gallery image not useful for recognition because the corresponding patch from the probe image is under disguise effect and matching a biometric patch with a non-biometric patch may lead to incorrect information.

$$\mathbf{F}^u(I^p, I^g) = \mathbf{F}(I^p) \wedge \mathbf{F}(I^g) \tag{2.3}$$

The patch classification algorithm explained in previous Section classifies the patches into biometric and non-biometric, and Eq. 2.3 provides information that *for a given gallery-probe pair, which patches should be used for face recognition*. Note that, in order to take advantage of patch classification, the face recognition approach has to be patch-based. Therefore, we propose to use LBP [21] which is one of the widely used patch-based descriptors for face recognition. If $\mathrm{desc}^{\mathrm{I}}_{\mathrm{ij}}$ represents the LBP descriptor of ij patch of image I, and the χ^2-distance between two LBP descriptors is represented as $\mathrm{dist}(\cdot,\cdot)$, then the distance D_{I^p,I^g} between two images, I^p and I^g, is calculated as:

$$\begin{aligned} D_{I^p,I^g} &= \tfrac{1}{\eta} \sum_i \sum_j \mathrm{dist}(\mathrm{desc}^{I^p}_{ij}, \mathrm{desc}^{I^g}_{ij}) \mathbf{F}^u(I^p, I^g)_{ij} \\ &\text{where } \eta = \sum_i \sum_j \mathbf{F}^u(I^p, I^g)_{ij} \end{aligned} \tag{2.4}$$

and $\mathbf{F}^u(I^p, I^g)_{ij}$ is obtained using Eq. 2.3.

2.2.3 Results of the Proposed Algorithm

This section demonstrates the results of the proposed face recognition framework which includes the patch classification algorithm and LBP based face recognition, along with its comparison to Sparse Representation Classifier (SRC) and a commercial off-the-shelf system (COTS). We also compare the results of proposed algorithm with the results of human evaluation results.

All the images in the database are divided into 5×5 non-overlapping rectangular patches of size 26×30 pixels. Every patch is manually annotated as biometric or non-biometric to create the ground truth for training as well as evaluation. If more than half of the patch is covered with accessories, it is annotated as a non-biometric patch. Images of randomly chosen 35 subjects form the training set and the images from the remaining 40 subjects are used for testing. The training set thus contains 8050 patches (322 images×25 patches), out of which 6324 patches are biometric and 1726 patches are non-biometric. Similarly, the testing set comprises 8975 patches (359 images×25 patches) amongst which 6944 are biometric and 2031 are non-biometric. Depending on the disguise accessories used, the number of biometric patches in every image vary. Figure 2-7 shows the distribution of (annotated) biometric patches in the training and testing splits. The distribution provides an overview of disguise characteristics of the train and test sets. For example, it shows that in the training set there are about 180 images with no disguise (number of biometric patches = 25) and both the sets contains very minute number of images that has almost whole image under disguise ($0 \geq$ number of biometric patches ≥ 5).

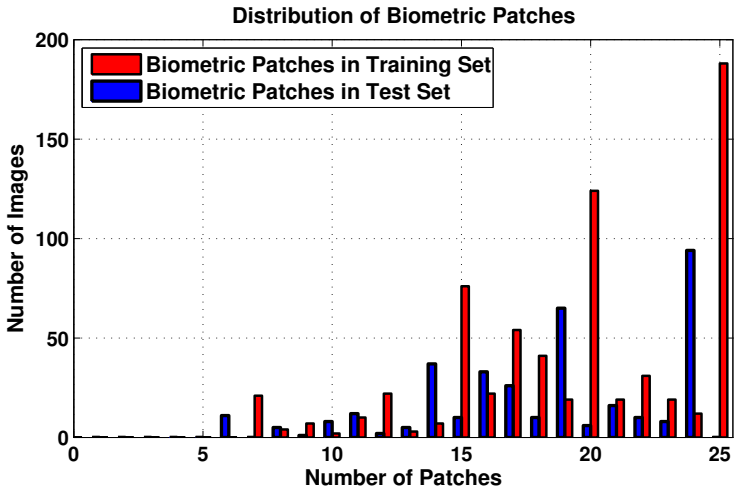


Figure 2-7: The distribution of biometric patches in the training and test sets.

Patch Classification using ITE

As explained earlier, for each patch, the ITE features are computed using Eq. 2.1; and min-max normalization is performed to map the values in the interval $[-1, 1]$. The normalized descriptor is provided as input to SVM with Radial Basis Function kernel for patch classification. The ker-

nel parameter and misclassification cost are estimated using grid search along with 5-fold cross validation. In grid search, parameters that provide the maximum training accuracy are considered as optimum. Since ITE is a concatenation of LBP and intensity values, the efficacy of ITE is compared with LBP and pixel intensity values. LBP histograms, intensity histograms, and ITE histograms are computed and provided as input to SVM separately for classification. Receiver Operating Characteristics (ROC) curves for patch classification representing the results of these experiments are shown in Figure 2-8. Note that, ITE provides better results compared to either texture or intensity information for patch classification. This supports our hypothesis that *concatenation of texture and intensity features should yield better patch classification results*.

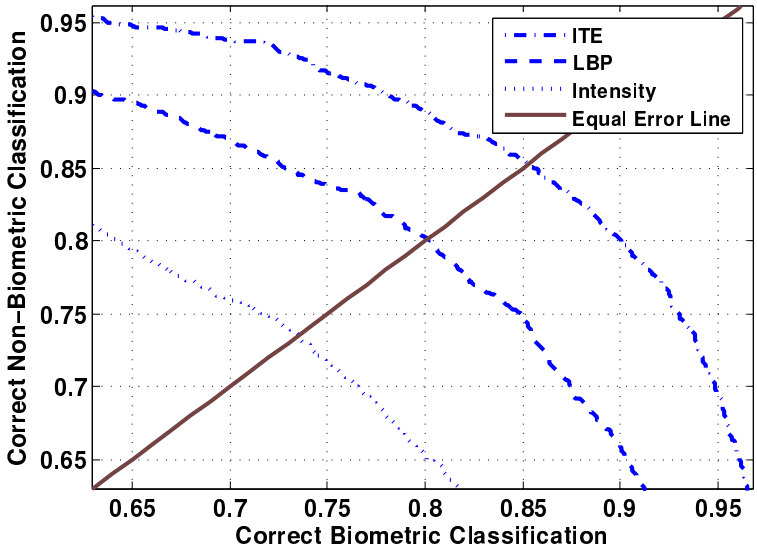


Figure 2-8: ROC curves for patch classification.

Performance Evaluation of Anāvṛta

The output of patch classification yields biometric patches which are used for feature extraction and matching. For evaluating the proposed face matching approach, the testing set is divided into two parts: gallery and probe. For each subject, one neutral face image, and four other randomly selected images are taken as gallery and the remaining images constitute the probe/query set. Hence, there are total 200 gallery images and 159 probe images. We have performed experiments with five random cross validation trials. The experiments are performed in verification mode and the results are reported in terms of ROC curve and verification accuracy at 0.1%, 1.0% and 10% False Accept

Rate (FAR). To understand the importance and effectiveness of performing patch classification, we performed the following three experiments.

1. Face recognition with biometric patches is classified using ITE and SVM classifier,
2. Face recognition with manually annotated biometric patches, and
3. Face recognition with all the patches (normal LBP approach without any patch removal)

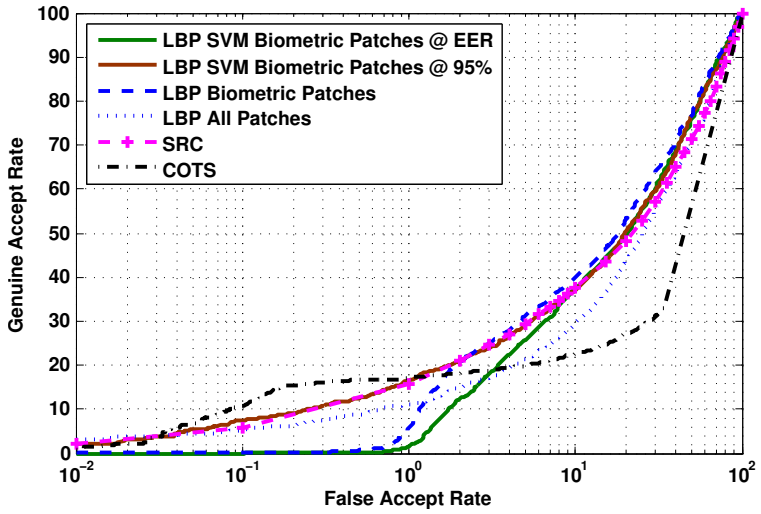


Figure 2-9: The results of the proposed face recognition framework using LBP descriptor.

The results of face recognition are shown in Figure 2-9. For FAR>1%, using only ground truth biometric patches results in better accuracy than using all the patches for face recognition. The performance of the proposed framework depends significantly on the performance of the patch classification algorithm. Intuitively, rejecting a non-biometric patch is less benefitting than the loss incurred by wrongly rejecting a biometric patch. From the ROC curve of patch classification shown in Figure 2-8, it can be analyzed that at equal error rate (EER), 15% of the biometric patches are being misclassified. show that the performance of face recognition reduces when the threshold of patch classification is chosen at EER. The ROC curves in Figure 2-9 show that the performance of face recognition reduces when the threshold of patch classification is chosen at EER. This may be attributed to the reduction in the number of biometric patches used for face recognition at that threshold. However, for 95% correct biometric patch classification (Figure 2-8), even though the number of correctly classified non-biometric patches decreases, the face recognition algorithm is

receiving more biometric patches as input and the proposed face recognition framework yields better performance than simple LBP based approach. This supports our hypothesis that not using non-biometric patches for recognition can result in better accuracy.

Comparison with COTS and Sparse Representation

In this section, we present a comparison with FaceVacs commercial off-the-shelf face recognition system (referred as COTS) and sparse representation classifier (SRC) [27]. Note that, SRC is a recent technique in literature for addressing occlusion/disguise. In SRC, the residual is considered as the dissimilarity measure of the gallery-probe pair. For evaluating the performance of the proposed framework, we have utilized all the gallery and probe images irrespective of the information content or image quality. However, COTS used in this research has inbuilt algorithms for quality assessment and enrollment. The thresholds for enrolling a gallery image are very strict whereas for probe images, it is relaxed. Out of the 200 gallery images, COTS enrolled approximately 60% of the gallery images and the remaining images were considered as *failure to enroll* whereas all the probe images were processed successfully. It is also observed that if the face image does not contain any non-biometric patch, then the probability of getting enrolled in the COTS is higher. However, for a fair comparison, we have overridden the COTS to include all 200 images in the gallery. Figure 2-9 and Table 2.6 demonstrate the results of COTS and SRC along with the proposed algorithm.

Table 2.6: Results from automated algorithms. Genuine accept rates and their standard deviations at different false accept rates of the proposed approach along with comparison to COTS and SRC.

Approach	Verification Accuracy @ FAR		
	0.1%	1.0%	10%
SRC	5.6 ± 1.3	15.5 ± 1.6	37.7 ± 1.8
COTS	10.9 ± 2.4	17.1 ± 1.5	22.5 ± 1.2
Proposed	7.4 ± 0.7	16.6 ± 0.5	38.1 ± 0.6

For face databases captured in constrained environment with cooperative users, face recognition algorithms yield high GAR, and it increases with increase in FAR [113]. However, this kind of trend is not found on this dataset with any of the three algorithms, thereby, showing the chal-

lenging nature of the database itself. It can be observed that COTS is not able to classify the faces under disguises very well as corresponding GAR does not increase much with increase in FAR. For lower FAR (<0.05%), all the approaches shown in comparison exhibit very poor performance. From approximately 0.2% till 5% FAR, the verification rate of COTS improves from 16% to 20% GAR. This may be attributed to COTS discarding many samples due to internal minimum quality criterion. For the same range of 0.2% to 5% FAR, the proposed approach yields up to 30% GAR. For almost whole range of FAR, the proposed approach is comparable to SRC. As shown in Table 2.6, although the performance reported by the proposed approach is not as high as it is usually reported in face recognition literature, it outperforms one of the state-of-art commercial systems and is comparable with a widely used technique (i.e. sparse representation).

In the evaluation of the proposed algorithm, it is observed that the performance of local (patch-based) face recognition algorithm can be improved by rejecting the face patches that contain disguise . Strict rejection of non-biometric patches leads to lower GAR at lower FAR. However, as discussed earlier a flexible patch classification at 95% correct biometric patch classification exhibits higher GAR even at lower FAR. Moreover, for FAR$>$ 1% the proposed automated algorithm outperforms the COTS which ends up rejecting large number of disguised face images which do not match its minimum criteria for processing. Although, the proposed algorithm equates to SRC [27] and outperforms COTS, the overall performance of $\sim$17% GAR at 1% FAR compared $90\%GAR@FAR = 1\%$ with very high accuracy that is usually reported for face verification of frontal non-disguised faces [113], suggest that significant amount of research is required to efficiently mitigate the effect of disguise variations.

2.2.4 Comparison of Human Responses with Automated Algorithms

As opposed to automated algorithm where for every image pair a match score is computed and compared with decision threshold to estimate the accuracy, human evaluation directly records their final decision. Therefore, for the automated algorithm ROC can be drawn by varying the threshold, whereas only a point (FAR-GAR pair) can be obtained on ROC from the human evaluation. Figure 2-10 represents the performance of all four Sets along with respective ROCs of the proposed automated algorithm and COTS. The key observations are as follows.

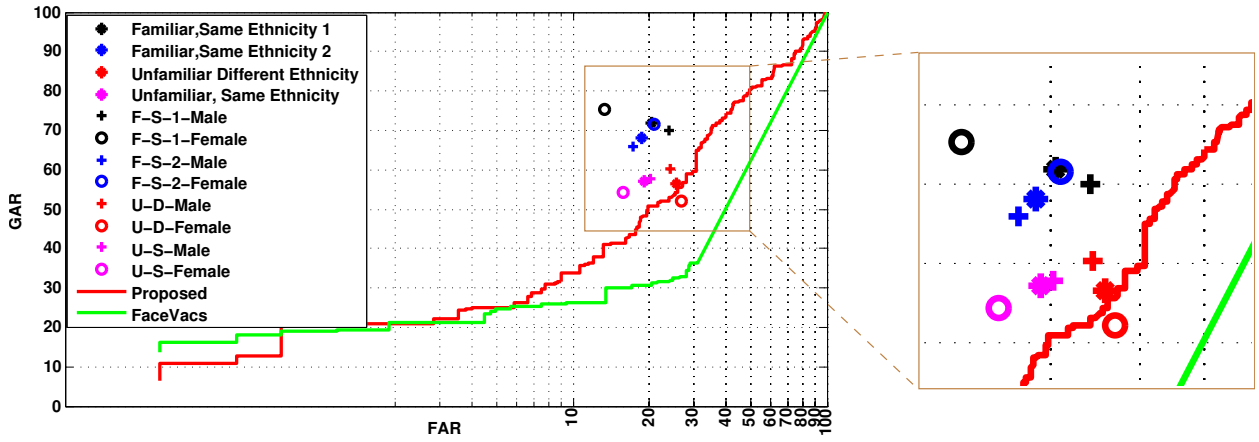


Figure 2-10: Performance of disguised face recognition by humans, with respect to familiarity and ethnicity. Analyzing the effect of familiarity and ethnicity on the performance of disguised face recognition by humans.

- The performance of Set FS (familiar, same ethnicity) is better than the one reported with automated algorithms (proposed and COTS).

- The ROC curve of the proposed algorithm passes through the performance point pertaining to Set UD. This is probably due to the fact that the automated algorithm does not encode familiarity or ethnicity, leading to no performance bias because of these two factors. Thus, proposed automated algorithm is comparable to humans recognizing unfamiliar subjects of different ethnicity. O'Toole *et al.* [102] have also observed that difference between the performance of humans and state-of-the-art face recognition algorithms were analogous to differences between humans recognizing familiar versus unfamiliar subjects. Researchers have also suggested that mental representation of familiar faces [114] helps make the familiar face recognition efficient compared to unfamiliar face recognition. If the machine counterpart of the mental representation is not incorporated somehow, the algorithms would face challenges similar to that of unfamiliar face recognition by humans.

- Although, FAR from human evaluations are smaller than that from automated algorithm, human performances exhibit considerably higher FARs ranging from 10%-30%.

- The proposed approach is a local approach and does not encode the holistic facial features whereas humans have access to both local and holistic facial information. Note that, we are

using the local approach because the holistic features can be corrupted by local disguises. The proposed local approach (ITE based patch classification+LBP based recognition) does improve performance over traditional local approach (LBP based recognition). However, the improved performance is only equivalent to the worst of human performance (Set UD) which favorably underlines the likely use of holistic facial features by humans. Therefore, simultaneous use of holistic and local facial features can lead to superior disguised face recognition performance.

- Our study on human evaluation suggests that ethnicity and familiarity of faces can greatly affect the face recognition performance. incorporating this information in face recognition algorithms can also provide improved matching accuracy.

2.3 Summary

This research presents a study on the effect of ethnicity and familiarity on the performance of face recognition in presence of disguise variations. The recognition accuracy of familiar-and-same-ethnicity subjects is found to be significantly better than that of unfamiliar-and-different-ethnicity. It is observed that if the ethnicity is same; unfamiliarity does not significantly affect correct rejection. Our experiments do not show any evidence of decrease in cross-ethnicity face recognition under disguise. We also observe that use of similar disguise accessories account for considerably high error rates.

Encoding the understanding from human evaluation, we propose an automated face recognition algorithm. The proposed algorithm consists of the ITE based patch classification (in biometric/non-biometric classes) and LBP based face recognition applied on classified biometric patches. The performance is evaluated on the IIIT-Delhi disguise database pertaining to 75 subjects. The proposed algorithm outperforms a COTS and classical LBP based face recognition. The performance of the proposed algorithm is comparable with SRC and the human performance of unfamiliar-and-different-ethnicity. Though we report performance improvement with the proposed algorithm, it is still an open research problem. The results of automatic algorithms are similar to unfamiliar face recognition performance of humans and therefore there is a scope for extending this research in the direction of both cognitive as well as automatic face recognition.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Heterogeneous Discriminant Analysis for Cross-View Face Recognition

With increasing focus on security and surveillance, face biometrics has found several new applications and challenges in real world scenarios. In terms of the current practices by law enforcement agencies, the legacy mugshot databases are captured with good quality face cameras operating in the visible spectrum (VIS) with inter-eye distance of at least 90 pixels [115]. However, for security and law enforcement applications, it is difficult to meet these standard requirements. For instance, in surveillance environment, when the illumination is not sufficient, majority of the surveillance cameras capture videos in the near infrared spectrum (NIR). Even in day-time environment, an image captured at a distance may have only 16×16 facial region for processing. For these applications, the corresponding gallery or database image is generally a good quality mugshot image captured in controlled environments. This leads to the challenge of *heterogeneity* in gallery and probe images. Fig. 3-1 shows samples of these heterogeneous face matching cases. This figure also showcases another interesting forensic and law enforcement application of matching composite sketch images with digital face images. In this problem, composite sketches are generated using a software tool based on eye-witness description and this synthetic sketch image is then matched against a database of mugshot face images. Since the information content in sketches and photos is different, matching them can be viewed as heterogeneous face matching problem.

The challenge of heterogeneous face recognition is posed by the fact that the *view*[1] of the

[1]The terms *view* and *domain/modality* are used synonymously in the heterogeneous face recognition literature.

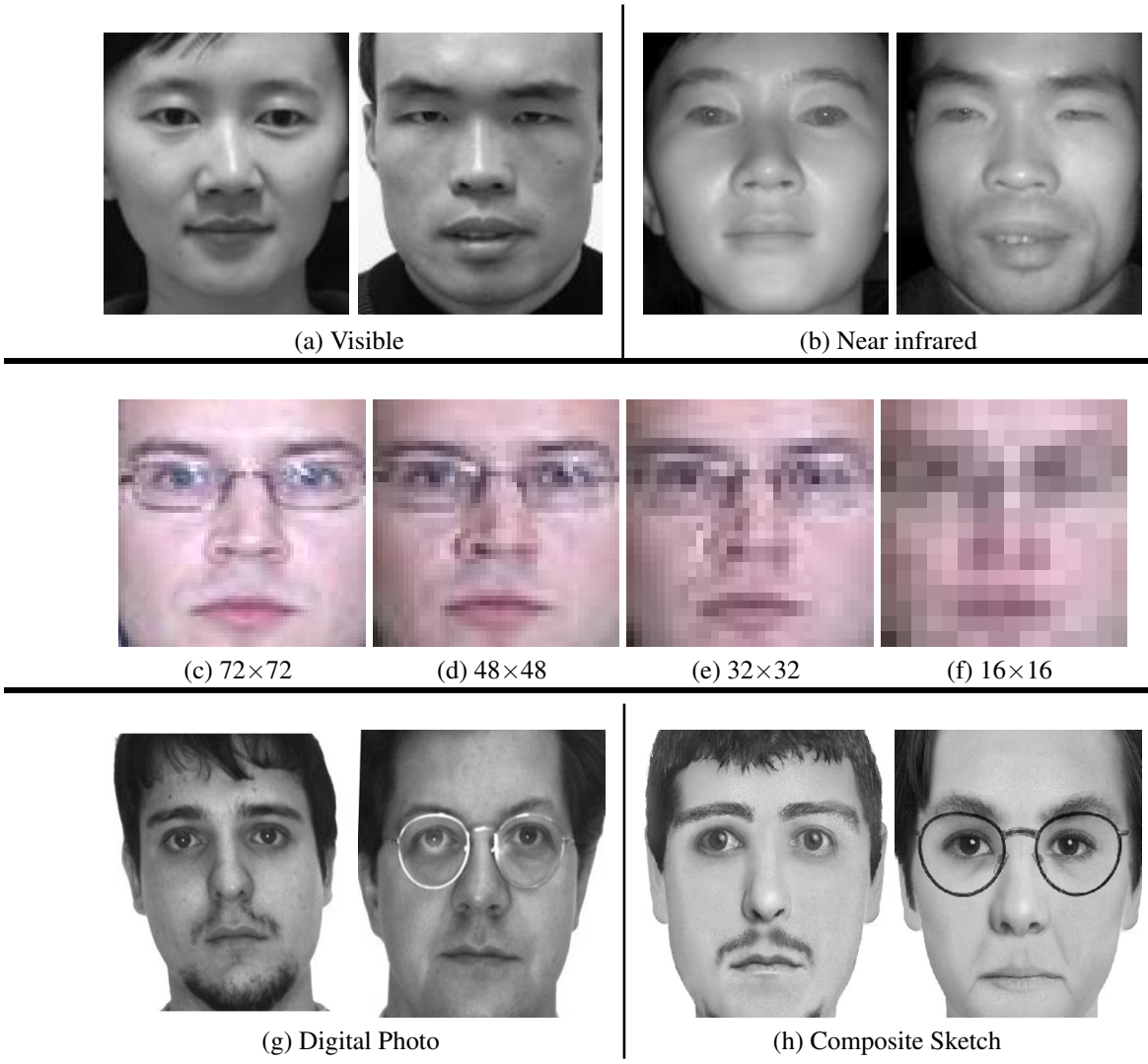


Figure 3-1: Examples of heterogeneous face recognition scenarios. Top row (a) & (b) shows heterogeneity due to spectrum difference. The middle row (c)-(f) illustrates heterogeneity due to resolution differences. (The images of different resolution are stretched to common sizes.) The bottom row shows (g)-(h) shows photo and composite sketches of the two subjects.

query face image is not same as that of the gallery image. In a broader sense, two face images are said to have *different views* if the facial information in the images is represented differently. For example, as shown in Fig. 3-2, visible and near infrared images are two views. The difference in views may arise due to several factors such as difference in sensors, their operating spectrum range, and difference in the process of sample generation. Most of the traditional face recognition research has focused on homogeneous matching [116], i.e., when both gallery and probe images have the same views. In recent past, researchers have addressed the challenges of heterogeneous

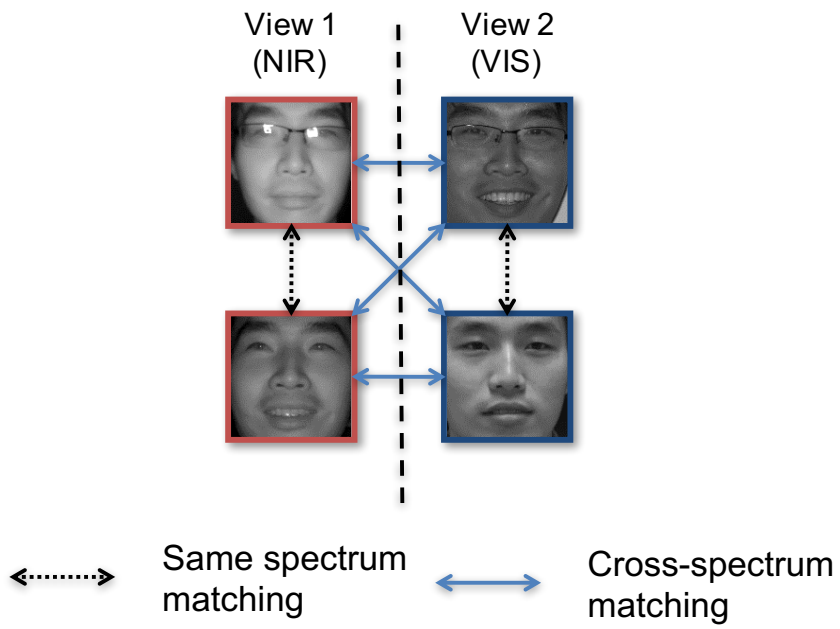


Figure 3-2: An example illustrating heterogeneous and homogeneous matching. Here, two views pertaining to spectrums (VIS and NIR) are shown. The solid lines represent comparisons corresponding to heterogeneous matching.

face recognition [117]–[122]. Compared to homogeneous face recognition, matching face images with different views is a challenging problem as heterogeneity leads to increase in the intraclass variability. Other widely explored covariates of pose, illumination, and expression can also cause increased intraclass variability and heterogeneity of views. For example, comparing a profile face image and frontal pose face image is a heterogeneous face matching problem. However, specific techniques for addressing these traditional covariates are explored widely; therefore, we focus on somewhat recent challenges.

The literature pertaining to heterogeneous face recognition can be grouped into two broad categories: 1) *heterogeneity invariant features* and 2) *heterogeneity aware classifiers*. Heterogeneity invariant feature based approaches focus on extracting features which are invariant across different views. The prominent research includes use of hand-crafted features such as variants of histogram of oriented gradient (HOG), Gabor, Weber, and local binary patterns (LBP) [123]–[127] and various learning-based features [128]–[131]. Heterogeneity aware classifier based approaches focus on learning a model using samples from both the views. In this research, we primarily focus on designing a *heterogeneity aware classifier*. Table 3.1 summarizes some of the research direc-

tions in emphasizing the role of classifiers to address cross-view variability of heterogeneous face recognition.

In one of the earliest research related to visible to near infrared matching, Yi *et al.* [117] proposed utilizing canonical correlation analysis (CCA) which finds the projections in an unsupervised manner. It computes two projection directions, one for each view such that the correlation between them is maximized in the projection space. Each spectrum can be considered as one view of data and CCA requires that the number of samples in both the views should be exactly same. Closely related to CCA, Sharma *et al.* [141] proposed generalized multi-view analysis (GMA) by adding a constraint that the multi-view samples of each class are as much close as possible. Recently, several other extensions of CCA to multi-view scenarios are also proposed [153]–[155]. Moreover, dictionary learning based multi-view extensions are also proposed [148], [156].

Lin and Tang [132] proposed common discriminant feature extractor (CDFE). The objective of CDFE is to learn one transformation function for each view, such that the empirical separability and the local consistency are maintained. Lie and Li [118] proposed coupled spectral regression (CSR) which also aims at obtaining two projection directions for each view. The objective is to obtain projection directions such that the samples from one view can be approximated by the projection of the corresponding sample in the second view, and vice versa. The authors also proposed kernel version of the same approach. Although the approach utilizes the correspondence between samples from both views, it does not use the class labels explicitly. Lie *et al.* [119], [137] further improved the CSR by allowing the samples from one view to contribute in finding the projection direction for the other view.

Li *et al.* [134] proposed to learn projection directions such that 1) the projections of the corresponding samples in both views should be similar and 2) the projections of a sample in one view should be also similar to the projections of the local neighbours of the corresponding sample. Klare *et al.* [120] proposed a prototyping based approach. It explores the intuition that across different views the relative coordinates of samples should remain similar. Therefore, the vector of similarities between the query sample and prototype samples in the corresponding view may be used as the feature. To facilitate non-linear classification these distances are measured in a kernel space using a discriminative learning algorithm. Biswas *et al.* [138], [139] proposed a multidimensional scaling (MDS) based approach for matching low resolution face images. The algorithm learns an

Table 3.1: Summary of selected research papers on heterogeneous face matching. VS=viewed sketch, FS=forensic sketch, CR=cross resolution, HR=high resolution, LR=low resolution, TH=thermal, VIS=visible, and NIR=near infrared.

Technique	**Approach/Objective function involves**	**Application**
CDFE [132] (2006)	empirical separability, local consistency	VIS-IR, Digital Photo - Sketch
CCA [117] (2007)	inter-view cross-correlation	VIS-NIR [133]
CSR [118] (2009)	sample reconstruction correspondence	VIS-NIR [133]
CLPM [134] (2010)	inter-view projection similarity, inter-view locality constraint	HR-LR
PLS [135] (2011)	maximization of covariance between latent scores	Pose Variations, HR - LR Digital Photo-Sketch [136])
ICSR, CDA [119], [137] (2012)	sample reconstruction, locality constraint	VIS-NIR [133], HR-LR, Digital Photo - Video Frame
MDS [138], [139] (2012, 2013)	learning manifold of one view constrained by pair-wise distances of other view	LR face matching
MvDA [140] (2012, 2016)	bringing same class samples (across views) closer and class means further in projected space	Pose Variations, VIS-NIR [133], Digital Photo - Sketch [136]
GMLDA, GMMFA [141] (2012)	optimizing between and within-class scatter in projected space and projecting same class samples near-by	Pose and Illumination Variations, Text - Image Retrieval
Kernel Prototype [120] (2013)	distance from prototypes as signature	VIS-NIR [133], VIS-TH, VIS-VS, VIS-FS
MMCM [142] (2013)	maximum-margin criterion	HR-LR face, HR-LR ocular
CFDA [143] (2014)	correlation of probability distributions	VIS-NIR [133]
THFM [144] (2014)	transductive learning	VIS-NIR [133], [145]
CDFL [121] (2015)	discriminative feature learning	VIS-NIR [146], VIS-VS [136], VIS-3D [57]
C-CBFD (+LDA) [147] (2015)	leaning compact, energy preserving, and evenly distributed binary coding such that encoding is correlated for sample of multiple views	VIS-NIR [146]
Joint dictionary [148] (2015)	learning a dictionary for analysis-by-synthesis based matching	VIS-NIR [146]
Shared representation [128] (2015)	Shared representation learning using Restricted Boltzmann machine	VIS-NIR [133], [146], Digital Photo-Sketch [136]
Reale *et al.* [129] (2016)	deep convolution neural network based	VIS-NIR [133], [146]
Frankenstein [130] (2016)	synthesis approach to train CNN models with limited training data	VIS-NIR [146]
TRIVET [131] (2016)	fine tuning of homogeneous model using heterogeneous dataset	VIS-NIR [146]
Jin *et al.* [149] (2016)	extreme learning machine based approach to address heterogeneity	VIS-NIR [133], VIS-3D [57]
HJB [150] (2017)	heterogeneous joint Bayesian learning	VIS-NIR [133], [146], ID-Spot [150]
G-HFR [151] (2017)	coupled representation similarity metric and graphical representations	Multiple databases including VIS-NIR [146]

MDS transformation which maps pairwise distances in kernel space of one view to corresponding pairwise distances of the other view. Siena *et al.* [142] proposed a maximum margin based approach. Its objective function is to find projection directions such that distances between the projections of samples of match pairs are minimized; and that such distance for match pair should be smaller than the same for non-match pairs. Recently, Li *et al.* [143] proposed a learning based feature descriptor in a two-level matching framework. Zhu *et al.* [144] proposed a transductive learning based framework which does not require to have face images of both the views (spectrums) for all the subjects in training. Kang *et al.* [157] focused on recognizing faces with multiple heterogeneous variations such as spectrum and distance/resolution. With advances in deep learning based approaches and their effectiveness in face recognition, researchers have explored these approaches for heterogeneous face recognition also. Some of the important research work in this direction include [128]–[131].

This research aims at making both theoretical and application oriented contributions for heterogeneous face matching. The key contributions are:

- Proposing a heterogeneity-aware classifier, termed as Heterogeneous Discriminant Analysis (HDA), and its non-linear extension termed as kernel HDA (KHDA). These subspace based classifiers aim at reducing the inter-view intra-class variability and increasing the inter-view inter-class variability for heterogeneous face recognition.
- Presenting a heterogeneous face recognition algorithm designed using the proposed heterogeneity-aware subspace classifiers.
- The effectiveness of the proposed HDA and KHDA is demonstrated using multiple features on three challenging heterogeneous cross-view face recognition scenarios: (1) matching visible to near-infrared images, (2) matching cross-resolution face images, and (3) matching digital photo to composite sketch.

3.1 Heterogeneous Discriminant Analysis

To address the issue of heterogeneity in face recognition we propose a discriminant analysis based approach. In this context, the heterogeneity can arise due to factors such as spectrum variations as

shown in Fig. 3-2. The same individual may appear somewhat different in two different spectrums. It is our hypothesis that incorporating cross-view (e.g. cross-spectrum) information along with face specific feature space can improve heterogeneous matching. The proposed heterogeneous discriminant analysis is inspired from the formulation of linear discriminant analysis. Therefore, we first briefly summarize the formulation and limitations of LDA followed by presenting the details of HDA.

Traditionally, intra-class and inter-class variabilities are represented using within-class (S_W) and between-class scatter matrices (S_B).

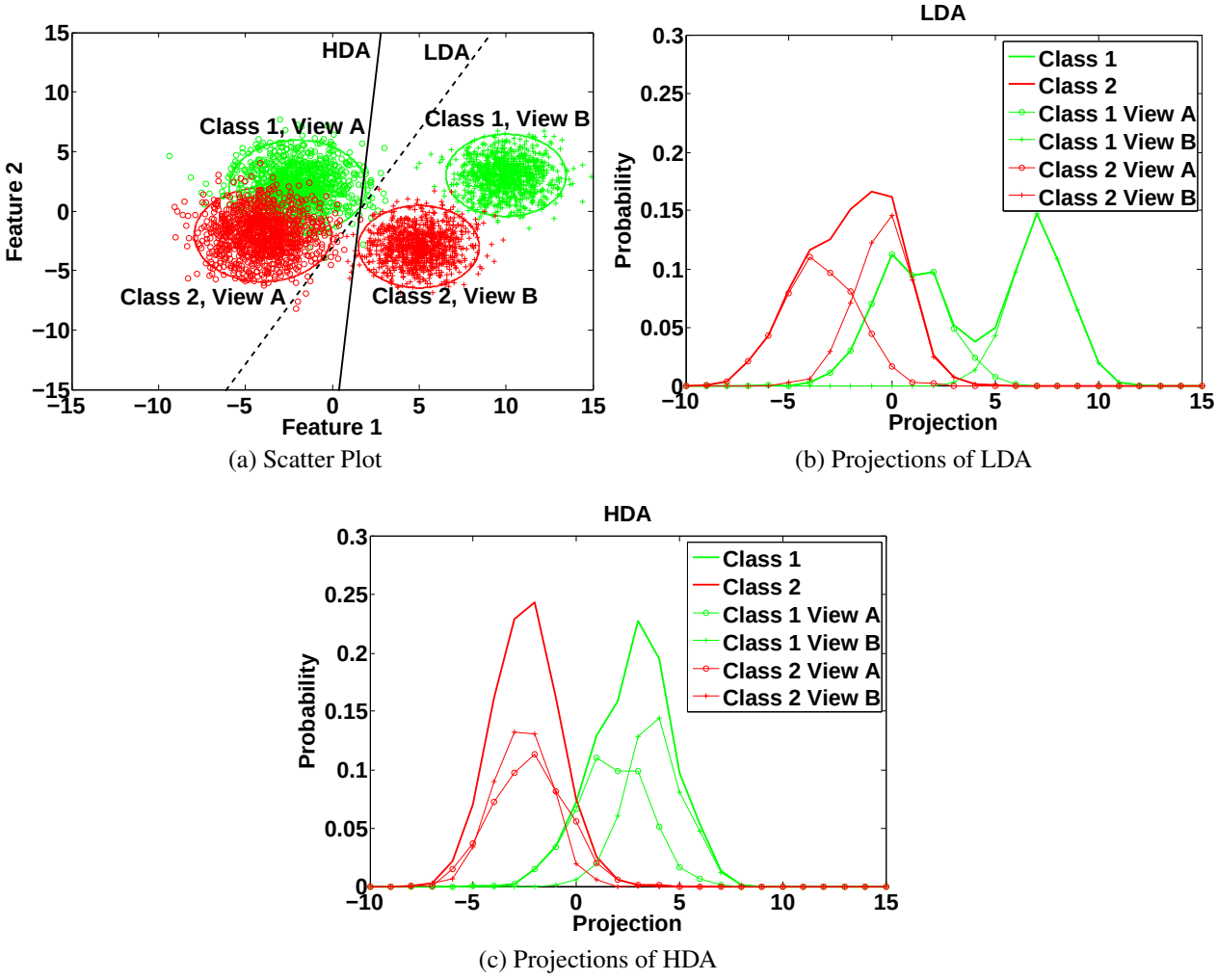
$$S_W = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (x_{i,j} - \mu_i)(x_{i,j} - \mu_i)^T, \quad S_B = \sum_{i=1}^{c} \sum_{l=i+1}^{c} (\mu_i - \mu_l)(\mu_i - \mu_l)^T \tag{3.1}$$

Here, c is the total number of classes, $x_{i,j}$ represents the j^{th} sample of the i^{th} class, and μ_i is the mean of the i^{th} class. The Fisher criterion attempts to find the projection directions that minimize the intra-class variability and maximize the inter-class variability in the projected space.

$$J(w) = \frac{|w^T S_B w|}{|w^T S_W w|} \tag{3.2}$$

The way the scatter matrices are defined ensures that *all* the samples are as close to the corresponding class mean as possible, and that class means are as apart as possible. Any new sample resembling the samples of a certain class would get projected near the corresponding class mean. LDA attempts to optimize the projection directions assuming that the data conforms to a normal distribution. Obtaining such a projection space is useful when the samples to be compared are homogeneous, i.e. there is no inherent difference in the sample representation. Even if we assume that each view of each class is normally distributed in itself, the restrictive constraint of LDA is not satisfied. As shown in Fig. 3-3a and Fig. 3-3b, when provided with a multi-view or heterogeneous data, the projection directions obtained from LDA may be suboptimal, and can affect the classification performance. Therefore, for heterogeneous matching problems, we propose to incorporate the view information while computing the between-class and within-class scatter matrices.

The formulation of the proposed Heterogeneous Discriminant Analysis is described in the following two stages: 1. adaptation of scatter matrices and 2. analytical solution.



(a) Scatter Plot

(b) Projections of LDA

(c) Projections of HDA

Figure 3-3: A toy example illustrating the effectiveness of HDA with multiple views. Class 1 and 2 are generated using Gaussian mixture of two modes resulting in two views. (a) represents the scatter plot and the projection directions obtained using LDA and HDA (without regularization). The histograms of projections of data samples on the LDA and HDA directions are shown in (b) and (c) respectively. Along with the overall class projection histograms, the histograms of projections of the constituting views are also shown (in plots with 'o' and '+' markers). The projection direction obtained by LDA is suboptimal, i.e. there are distinct view specific peaks, relatively more spread projection distributions, and there is a large inter-class overlap. On the other hand, HDA provides overlapping or very close view specific peaks, and less inter-class overlap.

3.1.1 Adaptation of scatter matrices

Let $x^a_{i,j}$ and $x^b_{i,j}$ denote the two views (A and B) of the j^{th} sample of the i^{th} class, respectively; and, n^a_i and n^b_i represent the number of samples in view A and B of the i^{th} class, respectively. $\chi^a_i = \{x^a_{i,j} | 1 \leq j \leq n^a_i\}$ represents the samples in view A of i^{th} class. For example, χ^a_i represents

the visible spectrum face images of i^{th} subject, and χ_i^b represents the near infrared spectrum face images of the subject.

- χ_1^a-χ_1^a and χ_1^a-χ_1^b are examples of match pairs i.e. face images in a pair belong to same subject.
- χ_1^a-χ_2^b and χ_1^a-χ_2^b are examples of non-match pairs consisting of face images of different subjects.
- χ_1^a-χ_1^a and χ_1^b-χ_2^b represent intra-view pairs where face images belong to same view.
- χ_1^a-χ_1^b and χ_1^b-χ_2^a are examples of inter-view pairs i.e. face images in a pair belong to different view.

As shown in Fig. 3-2, there can be four kinds of information: (i) inter-class intra-view difference, (ii) inter-class inter-view difference, (iii) intra-class intra-view difference, and (iv) intra-class inter-view difference. Optimizing the intra-view (homogeneous) distances would not contribute in achieving the goal of efficient heterogeneous matching. Therefore, the scatter matrices should be defined such that the objective function reduces the heterogeneity (inter-view variation) along with improving the classification accuracy. The distance between the *inter-view* samples of the *non-matching class* should be increased and the distance between *inter-view* samples of the *matching class* should be decreased. With this hypothesis, we propose the following two modifications in the scatter matrices for heterogeneous matching:

Inter-class Inter-view Difference encodes the difference between different views of two individuals (e.g. χ_1^a-χ_2^b and χ_1^b-χ_2^a pairs). This can be incorporated in the between-class scatter matrix.

Intra-class Inter-view Difference encodes the difference between two different views of one person (e.g. χ_1^a-χ_1^b and χ_2^b-χ_2^a pairs). This can be incorporated in the within-class scatter matrix. (See Fig. 3-3)

Incorporating these yields a projection space in which same class samples from different views are drawn closer; thereby fine tuning the objective function for heterogeneous matching.

The heterogeneous between-class scatter matrix (S_{HB}) encodes the difference between differ-

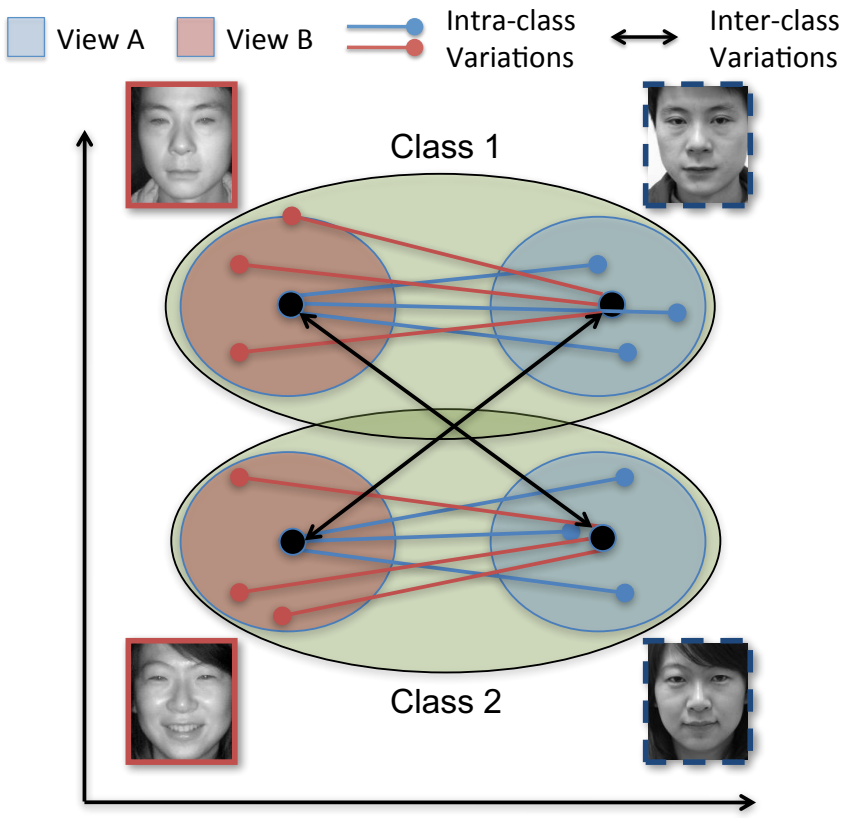


Figure 3-4: Graphical interpretation of the proposed HDA.

ent views of different classes.

$$S_{HB} = \sum_{i=1}^{c} \sum_{l=1, l\neq i}^{c} p_i^a p_l^b (\mu_i^a - \mu_l^b)(\mu_i^a - \mu_l^b)^T \tag{3.3}$$

$$\mu_i^a = \frac{1}{n_i^a} \sum_j x_{i,j}^a, \quad p_i^a = \frac{n_i^a}{n^a + n^b}, \quad \mu_i^b = \frac{1}{n_i^b} \sum_j x_{i,j}^b, \quad p_i^b = \frac{n_i^b}{n^a + n^b}$$

Here, μ_i^a and p_i^a are the mean and prior of view A of class i, respectively; n^a represents the number of samples in view A. Similarly, μ_i^b and p_i^b represent the mean and prior of view B of class i, respectively; n^b represents the number of samples in view B. n_i^a and n_i^b represent the number of samples in view A and B of the i^{th} class, respectively and n^c represents the total number of classes. Note that, unlike CCA, number of samples does not have to be equal in both views. The

within-class scatter matrix S_{HW} is proposed as

$$S_{HW} = \sum_{i=1}^{c} \left(\frac{1}{n_i^a} \sum_{j=1}^{n_i^a} (x_{i,j}^a - \mu_i^b)(x_{i,j}^a - \mu_i^b)^T + \frac{1}{n_i^b} \sum_{j=1}^{n_i^b} (x_{i,j}^b - \mu_i^a)(x_{i,j}^b - \mu_i^a)^T \right) \tag{3.4}$$

Since the proposed technique encodes data heterogeneity in the objective function and utilizes the definitions of between-class and within-class scatter matrices, it is termed as heterogeneous discriminant analysis. Following the Fisher criterion, the optimization criteria of HDA is proposed as,

$$w = \arg\max_w J(w) = \arg\max_w \frac{\left| w^T S_{HB} w \right|}{\left| w^T S_{HW} w \right|} \tag{3.5}$$

The optimization problem in Eq. 3.5 is modeled as a generalized eigenvalue decomposition problem; which results into a closed form solution such that w is the set of top eigenvectors of $S_{HW}^{-1} S_{HB}$. The geometric interpretation of HDA in Fig. 3-4 shows that the objective function in Eq. 3.5 tries to achieve the following in the projected space: 1) Bring samples χ_1^a closer to mean μ_1^b of χ_1^b and vice versa; and similarly for class 2. This reduces the inter-view distance within each class, e.g. the projections of visible and NIR images of the same person becomes similar. 2) Increase the distance between mean μ_1^a of χ_1^a and mean μ_2^b of χ_2^b; and similarly increase the distance between mean of χ_1^b and mean of χ_2^a, i.e. the projections of mean visible face image of a subject becomes different from the mean NIR face image of another subject. The proposed way of encoding the inter-class (Eq. 3.3) and intra-class (Eq. 3.4) variations in the heterogeneous scenario requires that both the views are of the same dimensionality. In the application domain of face recognition, this is usually not an unrealistic constraint as, in practice, same kind of features, with same dimensionality, are extracted from both the views [123].

The time complexity of computing S_{HB} and S_{HW} is $O(nd^2)$ and $O(c^2d^2)$, respectively. The generalized eigenvalue decomposition in Eq. 3.5 has time complexity of $O(d^3)$, where n, d, and c are the number of training samples, feature dimensionality, and number of classes, respectively.

In some applications including face recognition, the number of training samples is often limited. If the number of training samples is less than the feature dimensionality, it leads to problems such as singular within-class scatter matrix. In literature, it is also known as the small sample size

problem and a shrinkage regularization is generally used to address the issue [158]. Utilizing the shrinkage regularization Eq. 3.5 is updated as,

$$J(w) = \frac{\left|w^T S_{HB} w\right|}{\left|w^T \left((1-\lambda) S_{HW} + \lambda I\right) w\right|} \tag{3.6}$$

Here, I represents the identity matrix and λ is the regularization parameter. Note that $\lambda = 0$ results in no regularization, whereas $\lambda = 1$ results into not utilizing the within-class scatter matrix S_{HW}.

Analytical Solution of HDA

We further analyze the objective function in Eq. 3.5. Using the representer theorem [159], the projection direction in w can be written as linear sum of the samples, i.e.

$$w = \sum_{p=1}^{n^a} \alpha_p x_p^a + \sum_{q=1}^{n^b} \beta_q x_q^b \tag{3.7}$$

where, x_p^a is the p^{th} sample of view A, and α_p and β_q are their corresponding coefficients. In this formulation, the problem of finding w is converted into finding the coefficient vectors α and β. We begin by obtaining the expression for projection of a sample using coefficient based definition of projection direction w. The projection of j^{th} sample of i^{th} class from view A is given as

$$w^T x_{i,j}^a = \sum_{p=1}^{n^a} \alpha_p x_p^a \cdot x_{i,j}^a + \sum_{q=1}^{n^b} \beta_q x_q^b \cdot x_{i,j}^a = [\alpha^T \beta^T] \begin{bmatrix} X^a \\ X^b \end{bmatrix} x_{i,j}^a \tag{3.8}$$

Essentially, the projection of a sample is equal to weighted sum of its dot product with all the training samples. In other words, the projection of a sample is defined based on its structural arrangement with respect to samples from both the views. Similarly, the projection of the mean of i^{th} class of view A is as follows

$$w^T \mu_i^a = \sum_{p=1}^{n^a} \alpha_p x_p^a \cdot \mu_i^a + \sum_{q=1}^{n^b} \beta_q x_q^b \cdot \mu_i^a = \alpha^T \mathcal{M} \mathcal{A}_i^a + \beta^T \mathcal{M} \mathcal{B}_i^a \tag{3.9}$$

where $(\mathcal{MA}_i^a)_p = \frac{1}{n_i^a}\sum_{r=1}^{n_i^a} x_p^a \cdot x_{i,r}^a = x_p^a \cdot \mu_i^a$, and $(\mathcal{MB}_i^a)_q = \frac{1}{n_i^a}\sum_{s=1}^{n_i^a} x_q^b \cdot x_{i,s}^a = x_q^b \cdot \mu_i^a$

$$w^T\mu_i^a = [\alpha^T \beta^T] \begin{bmatrix} \mathcal{MA}_i^a \\ \mathcal{MB}_i^a \end{bmatrix} = [\alpha^T \beta^T] \begin{bmatrix} X^a \\ X^b \end{bmatrix} \mu_i^a \tag{3.10}$$

The derivation shows that the projection of mean μ_i^a in direction w can be obtained by taking its dot product with the samples. In a way, the vectors $\mathcal{MA}_i^a$ and $\mathcal{MB}_i^a$ encode the relative structural arrangement of mean μ_i^a with respect to the samples of views A and B respectively.

Using the samples and mean projections derived in Eq. 3.8 and Eq. 3.10 respectively, we formulate the $w^T S_{HB} w$ term of the optimization criteria shown in Eq. 3.5 as follows

$$w^T S_{HB} w = \sum_{i=1}^{c} \sum_{l=1, l\neq i}^{c} p_i^a p_l^b \left(w^T\mu_i^a - w^T\mu_l^b\right)\left(w^T\mu_i^a - w^T\mu_l^b\right)^T \tag{3.11}$$

Substituting Eq. 3.10 in Eq. 3.11 results in $w^T S_{HB} w = [\alpha^T \beta^T] M_* \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ where,

$$M_* = \sum_{i=1}^{c} \sum_{l=1, l\neq i}^{c} p_i^a p_l^b \begin{bmatrix} \mathcal{MA}_i^a - \mathcal{MA}_l^b \\ \mathcal{MB}_i^a - \mathcal{MB}_l^b \end{bmatrix} \begin{bmatrix} \mathcal{MA}_i^a - \mathcal{MA}_l^b \\ \mathcal{MB}_i^a - \mathcal{MB}_l^b \end{bmatrix}^T \tag{3.12}$$

The inter-view within-class variability in the projected space ($w^T S_{HW} w$) which is to be minimized can be rewritten as following with the help of Eq. 3.3.

$$\begin{aligned} w^T S_{HW} w = \sum_{i=1}^{c} \Bigg(& \frac{1}{n_i^a} \sum_{j=1}^{n_i^a} \left(w^T x_{i,j}^a - w^T\mu_i^b\right)\left(w^T x_{i,j}^a - w^T\mu_i^b\right)^T \\ & + \frac{1}{n_i^b} \sum_{j=1}^{n_i^b} (w^T x_{i,j}^b - w^T\mu_i^a)(w^T x_{i,j}^b - w^T\mu_i^a)^T \Bigg) \end{aligned} \tag{3.13}$$

Similarly, substituting Eq. 3.10 and Eq. 3.8 in Eq. 3.13 results in $w^T S_{HW}^{\Phi} w = [\alpha^T \beta^T] N_* \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$

where

$$N_* = \sum_{i=1}^{c} \left(\frac{1}{n_i^a} \sum_{j=1}^{n_i^a} \begin{bmatrix} \mathcal{KA}_{i,j}^a - \mathcal{MA}_i^b \\ \mathcal{KB}_{i,j}^a - \mathcal{MB}_i^b \end{bmatrix} \begin{bmatrix} \mathcal{KA}_{i,j}^a - \mathcal{MA}_i^b \\ \mathcal{KB}_{i,j}^a - \mathcal{MB}_i^b \end{bmatrix}^T + \frac{1}{n_i^b} \sum_{j=1}^{n_i^b} \begin{bmatrix} \mathcal{KA}_{i,j}^b - \mathcal{MA}_i^a \\ \mathcal{KB}_{i,j}^b - \mathcal{MB}_i^a \end{bmatrix} \begin{bmatrix} \mathcal{KA}_{i,j}^b - \mathcal{MA}_i^a \\ \mathcal{KB}_{i,j}^b - \mathcal{MB}_i^a \end{bmatrix}^T \right) \tag{3.14}$$

where, $(\mathcal{KA}_{i,j}^a)_p = x_p^a \cdot x_{i,j}^a$, $(\mathcal{KA}_{i,j}^b)_p = x_p^a \cdot x_{i,j}^b$, $(\mathcal{KB}_{i,j}^a)_q = x_q^b \cdot x_{i,j}^a$, and $(\mathcal{KB}_{i,j}^b)_q = x_q^b \cdot x_{i,j}^b$. Substituting Eq. 3.12 and Eq. 3.14 in the optimization function in Eq. 3.5 yields the following criterion

$$J(\alpha, \beta) = \left| [\alpha^T \beta^T] M_* \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right| \div \left| [\alpha^T \beta^T] N_* \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right| \tag{3.15}$$

Maximization of the criterion is modeled in terms of the generalized eigen decomposition problem. Thus, utilizing top $c-1$ eigen vectors of $N_*^{-1} M_*$ as coefficients maximizes the criterion function. In practice, N_* is often singular, therefore shrinkage regularization is utilized as follows:

$$J(\alpha, \beta) = \left| [\alpha^T \beta^T] M_* \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right| \div \left| [\alpha^T \beta^T] \left[(1-\lambda) N_* + \lambda \right] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right| \tag{3.16}$$

Note that, the criterion in Eq. 3.6 and Eq. 3.16 are different representations of the same optimization function, i.e. to minimize the inter-view intra-class variability and to maximize the inter-view inter-class variability in projected space. The matrix M_* and N_* are analogous to S_{HB} and S_{HW} respectively. However, M_* and N_* are $n \times n$ matrices and S_{HB} and S_{HW} are $d \times d$ matrices (d=feature dimensionality, n=number of samples). Each element of matrices M_* and N_* encodes the variability corresponding to a sample pair, whereas, the same for S_{HB} and S_{HW} encodes the variability corresponding to a feature pair. If $d < n$, the criterion in Eq. 3.6 is computationally more efficient than Eq. 3.16 but if $d > n$, Eq. 3.16 is computationally more efficient than Eq. 3.6.

The criterion in Eq. 3.16 has a specific advantage over Eq. 3.6; i.e. the matrices M_* and N_*

can be computed from the Gram matrix, and that the computation of M_* and N_* do not necessarily require knowledge of actual data samples. This is a crucial property to formulate kernel extension of HDA with the help of kernel trick [160].

Let $\Phi(\cdot)$ be a transformation function that projects data in a higher dimensional space. With the assumption that the data is linearly separable in a higher dimension space, an appropriate transformation $x \rightarrow \Phi(x)$ can yield a representation where HDA can yield better classification than in the input space of x. In the higher dimension space, the entries of Gram matrices become the dot products of the transformed data points, resulting in the following modification:

$$
\begin{aligned}
&(\mathcal{KA}^a_{i,j})_p = \Phi(x^a_p) \cdot \Phi(x^a_{i,j}), (\mathcal{KA}^b_{i,j})_p = \Phi(x^a_p) \cdot \Phi(x^b_{i,j}) \\
&(\mathcal{KB}^a_{i,j})_q = \Phi(x^b_q) \cdot \Phi(x^a_{i,j}), (\mathcal{KB}^b_{i,j})_q = \Phi(x^b_q) \cdot \Phi(x^b_{i,j}) \\
&(\mathcal{MA}^a_i)_p = \frac{1}{n^a_i} \sum_{s=1}^{n^a_i} \Phi(x)^a_p \cdot \Phi(x)^a_{i,s} \text{ and } (\mathcal{MB}^a_i)_q = \frac{1}{n^a_i} \sum_{s=1}^{n^a_i} \Phi^b_q \cdot \Phi(x)^a_{i,s}
\end{aligned} \tag{3.17}
$$

A kernel function can be defined as $k(x, y) = \Phi(x) \cdot \Phi(y)$ to facilitate the computations of the aforementioned matrices bypassing the data transformation stage. Any valid kernel function can be utilized for this purpose, e.g. radial basis function $\left(k(x, y) = \exp\left(\frac{|x-y|^2}{2t}\right)\right)$ and polynomial function $\left(k(x, y) = (1 + x \cdot y)^d\right)$. Eq. 3.17 forms the basis for applying HDA in higher dimensional space. This non-linear extension of HDA is termed as kernel HDA (KHDA). Intuitively, KHDA is expected to model the non-linearly separable classes more effectively as compared to HDA.

3.1.2 Visualization

To visualize the functioning of the proposed HDA as opposed to LDA, the distributions of the projections obtained using LDA and HDA are shown in Fig. 3-3b and Fig. 3-3c respectively. Table 3.2 presents a quantitative analysis in terms of the overlap between projections of views of both classes. The overlap between two histograms is calculated as $\sum_m \min(h_1(m), h_2(m))$, where $h_1(m)$ and $h_2(m)$ are the values of the m^{th} bin of the first and second histograms respectively. In the ideal case, the projections of different views of the same class should completely overlap (i.e. area of overlap 0.5) and the projections of the views of different classes should be non-overlapping

Table 3.2: Analyzing the overlap of projection distributions in Fig. 3-3b and Fig. 3-3c.

Pair	Overlap		
	Ideal	**LDA**	**HDA**
Overall			
class 1 - class 2	0.000	**0.356**	0.159
Homogeneous			
view A class 1 - view A class 2	0.000	0.110	0.135
view B class 1 - view B class 2	0.000	0.005	0.013
Heterogeneous			
view A class 1 - view B class 2	0.000	**0.351**	0.076
view A class 2 - view B class 1	0.000	0.000	0.034
view A class 1 - view B class 1	0.500	0.025	0.261
view A class 2 - view B class 2	0.500	0.174	0.429

(i.e. area of overlap 0). Since LDA does not take into account the view information, the overlap between projections of both classes is large. Further, it is interesting to note that LDA yields a significant overlap of 0.351 between view A of class 1 and view B of class 2. Such overlap can deteriorate the heterogeneous matching performance. In the heterogeneous analysis (last two rows of Table 3.2), the overlap between projections of two views of the same class is relatively low. Note that view A and view B of class 1 result in two individual peaks. This also increases the intra-class variation, i.e. projection distributions of both classes are *spread* rather than *peaked*.

HDA yields better projection directions with less than 50% of inter-class overlap compared to LDA. For the homogeneous matching scenarios (fourth and fifth row), HDA has marginally poor overlap compared to LDA. However, for the heterogeneous scenarios, the overlap of HDA is significantly lower for non-match pairs (seventh and eighth row), and higher for match pairs (last two rows).

3.2 Proposed Face Recognition Approach

The main objective of this research is to utilize the proposed heterogeneity-aware classifiers in conjunction with robust and unique features for heterogeneous face recognition. Fig. 3-5 showcases the steps involved in the face recognition pipeline. From the given input image, face region

Table 3.3: Summary of the datasets utilized for evaluating the proposed HDA and KHDA on two heterogeneous face recognition challenges.

Case Study	Gallery	Probe	Dataset	#Subject	#Images
Cross spectral	VIS	NIR	CASIA NIR-VIS-2.0 [146]	725	17,850
Cross resolution	HR	LR	CMU-MultiPIE [61]	337	18,420
Photo-to-Sketch	DP	CS	e-PRIP composite sketch [161], [162]	123	246

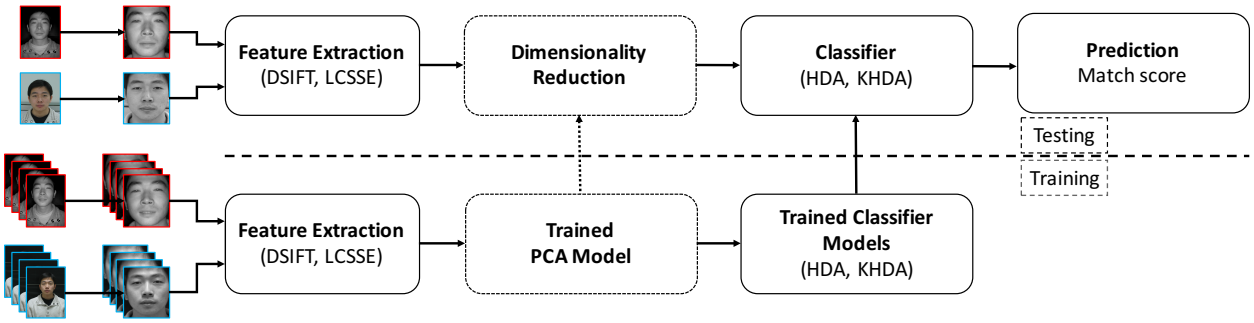


Figure 3-5: Steps involved in the face recognition pipeline with the proposed HDA and KHDA.

is detected using Haar face detector or manually annotated eye coordinates. It is our assertion that the proposed HDA and KHDA should yield good results with both handcrafted and learnt representations. Therefore, in this research, we have demonstrated the results with both sets of features.

In literature, it has been observed that Histogram of Oriented Gradient (HOG) and Local Binary Patterns (LBP) are commonly used handcrafted features for heterogeneous face matching [120], [163]. In 2014, Dhamecha *et al.* [123] compared the performance of different variants of HOG and showed that DSIFT yields the best results. Therefore, among handcrafted features, we have demonstrated the results with DSIFT[2] and LBP (uniform, $r = 1$, $p = 8$) [164] features.

For learnt representation, we use a recently proposed deep learning based feature extraction approach, termed as local class sparsity based supervised encoder (LCSSE) [165] that utilizes stacked auto-encoder [166] with $l_{2,1}$ regularization for supervision. LCSSE aims to learn the features with same sparsity signature across class, thus bringing the sparse representations of the same-class samples as close as possible in the feature space. The objective function of feature learning is to

[2]DSIFT features are extracted at keypoints on uniform grid and landmark points.

obtain encoding and decoding weights, $\mathbf{W}$ and $\mathbf{W}'$ such that the following objective function is minimized.

$$\arg\min_{\mathbf{W},\mathbf{W}'} ||\mathbf{X} - \mathbf{W}'\phi(\mathbf{W}\mathbf{X})||_F^2 + \lambda \sum_{c=1}^{C} ||\mathbf{W}\mathbf{X}_\mathbf{c}||_{2,1} \tag{3.18}$$

where, $\mathbf{X}$ is the training matrix, X_c represents matrix of training samples belonging to class $\mathbf{c}$, λ is the regularization parameter, C is the number of classes, $\phi(\cdot)$ is sigmoid function, and $||\cdot||_F$ and $||\cdot||_{2,1}$ represent Frobenius and $l_{2,1}$ norms, respectively. In [165], LCSSE based face recognition approach has shown state-of-the-art results on popular face databases such as LFW and PaSC. Therefore, LCSSE can be considered as a good choice for our experiments. In this research, we have used pre-trained LCSSE model and fine-tuned with the training samples for each case study.

As shown in Fig 3-5, once the features are obtained, they are projected on to a PCA space (preserving 99% eigenenergy) followed by projecting onto the HDA or KHDA space ($c-1$ dimensional). It is to be noted that learning of PCA subspace does not use class labels whereas HDA and KHDA training utilize identity labels and the view labels. While testing, the representation obtained for probe image is projected onto HDA or KHDA space, to make them better suited for heterogeneous matching. Finally, distance score between gallery and probe representation vectors x and y is computed using cosine distance measure i.e. $\left(1 - \frac{x \cdot y}{|x||y|}\right)$.

3.3 Experimental Evaluation

The effectiveness of the proposed heterogeneous discriminant algorithm is evaluated for three different case studies of heterogeneous face recognition: 1) visible to near infrared matching, 2) cross-resolution face matching, and 3) composite sketch (CS) to digital photo (DP) matching. For all three case studies, we have used publicly available benchmark databases: CASIA NIR-VIS-2.0 [146], CMU-MultiPIE [61] and e-PRIP composite sketch [161], [162]. Table 3.3 summarizes the characteristics of the three databases and Fig. 3-1 illustrates sample images from the databases. The experiments are performed with existing and published protocols so that the results can be directly compared with reported results.

Table 3.4: Rank-1 identification accuracies for visible to near infrared face matching on the CASIA NIR-VIS-2.0 database [146]. The experiments are performed by varying the feature extractors, classification models, and distance metrics.

Algorithm		DSIFT [168]	$LBP^{u2}_{8,1}$ [164]	LCSSE [165]
W/O DA	**Eucl**	12.6±0.9	4.9±0.7	50.3±8.3
	Cos	19.6±1.4	6.6±1.0	51.6±7.8
LDA	**Eucl**	56.7±2.2	17.7±2.2	82.3±4.8
	Cos	80.4±1.7	46.2±2.1	88.9±3.2
HDA	**Eucl**	58.0±2.1	26.9±2.1	**95.2±1.7**
	Cos	81.0±1.9	48.9±2.0	**96.8±0.9**

3.3.1 Cross Spectral Face Matching: Visible to NIR Images

Researchers have proposed several algorithms for VIS to NIR matching and primarily used the CASIA NIR-VIS-2.0 face dataset [146]. It consists of 17,850 NIR and VIS images (combined) pertaining to 725 subjects of varying age groups. The images are acquired in four different sessions. The protocol defined for performance evaluation consists of 10 splits of train and test sets for random subsampling cross validation. There are equal number of subjects in both train and test sets. As required by the predefined protocol, matching results are reported for both identification and verification scenarios. The identification performance is reported in terms of average rank-1 identification accuracy with standard deviation over 10 fold cross validation; and verification performance is reported in terms of GAR at 0.1% FAR.

The images are first detected and preprocessed. Seven landmarks (two eye corners, three points on nose, two lips corners) are detected [167] from the input face image and geometric normalization is applied to register the cropped face images. The output of preprocessing is grayscale face images of size 130×150 pixels. The aforementioned features (DSIFT, LBP, and LCSSE) are extracted from geometrically normalized face images. We evaluate the effectiveness of HDA over LDA. To compare the results with LDA, the pipeline shown in Fig. 3-5 is followed with the exception of using LDA instead of HDA. The results are reported in Table 3.4 and the key observations are discussed below.

Discriminative Learning using HDA: As shown in Table 3.4, without discriminant analysis (LDA

or HDA), the performance of individual features are significantly lower and deep learning based LCSSE yields around 50% rank-1 accuracy. The next experiment illustrates the effect of applying LDA on individual features. Table 3.4 shows that LDA improves the accuracy up to 60%. Comparing the performance of HDA with LDA shows that HDA outperforms LDA. Utilizing HDA in place of LDA for discriminative learning improves the results up to 12.9%. The improvement provided by HDA can be attributed to the fact that it learns, a discriminative subspace specifically for heterogeneous matching. Similar to the toy example shown in Fig. 3-3, it can be asserted that the multi-view information yields different clusters in the feature space. Under such scenarios, since the fundamental assumption of Gaussian data distribution is not satisfied, LDA can exhibit suboptimal results. However, by encoding the view label information, HDA is able to find better projection space, thereby yielding better results.

Effect of HDA on Features: Table 3.4 shows the performance obtained by applying HDA on three feature representations. The results show that the proposed HDA improves the accuracy of all three features by 40–60%. For instance, applying LCSSE with HDA improves the results by around 45%.

Direction vs Magnitude in Projection Space: For each of the classifier models and feature extractors, we have evaluated the performance of both Euclidean and cosine distance metrics. Cosine distance encodes only the difference in direction between samples, whereas the Euclidean distance encodes both direction and magnitude. As shown in Table 3.4, cosine distance generally yields higher accuracy over Euclidean distance. This shows that for heterogeneous matching, the magnitude of projections may not provide useful information, and only directional information can be used for matching.

Optimum Combination: From the above analysis, it can be seen that the proposed HDA in combination with DSIFT features and cosine distance measure yields significantly higher accuracy than LBP+HDA. Overall, utilizing HDA along with LCSSE features and cosine measure achieves the highest classification accuracy. Therefore, for the remaining experiments (and other case studies), we have demonstrated the results with DSIFT and LCSSE features and cosine distance measure along with proposed heterogeneity-aware classifiers.

Comparison with Existing Algorithms: We next compare the results of the proposed approaches

Table 3.5: Face recognition performance of the proposed and some existing algorithms for VIS to NIR face matching on CASIA NIR-VIS-2.0 dataset. †represents value obtained from ROC curve reported in the corresponding paper. *represents the results reported in [121], [147].

Algorithm	**Year**	**Rank-1 Accuracy (%)**	**GAR @ FAR=0.1%**
FaceVACS [123]	2014	58.6±1.2	52.9
Pixels as Features			
CCA [171]*	2004	28.5±3.4	10.8
PLS [135]*	2011	17.7±1.9	2.3
CDFE [132]*	2006	27.9±2.9	6.9
MvDA [140]*	2012	41.6±4.1	19.2
GMLDA [141]*	2012	23.7±1.4	5.1
GMMFA [141]*	2012	24.8±1.1	7.6
PCA+Symmetry+HCA [146]	2013	23.7±1.9	19.3
PIXEL+HDA	–	**41.4±1.3**	**31.4**
Other Features/Approaches			
DSIFT+SDA ($H = 2$)[172]	2006	75.7±1.9	54.8
Gabor + RBM + Remove 11 PCs [128]	2015	86.2±1.0	81.3
C-DFD (s=3) [173]*	2014	65.8±1.6	46.2
CDFL (s=3)[121]	2015	71.5±1.4	55.1
C-CBFD [147]	2015	56.6±2.4	20.4
C-CBFD+LDA [147]	2015	81.8±2.3	47.3
Joint Dictionary Learning [148]	2015	78.5±1.7	85.8
Saxena and Verbeek [174]	2016	85.9±0.9	78.0
Reale *et al.* [129]	2016	87.1±0.9	74.5
Frankenstein [130]	2016	85.1±0.8	-
TRIVET [131]	2016	95.7±0.5	91.0
Lezama *et al.* [175]	2016	89.6±0.9	-
MTC-ELM [176]	2016	89.1	-
Gabor+HJB [150]	2017	91.7±0.9	89.9
G-HFR [151]	2017	85.3±0.0	-
DSIFT+HDA	–	81.0±1.9	62.8
DSIFT+KHDA	–	83.1±1.7	62.1
LCSSE+HDA	–	**96.8±0.9**	**93.1**
LCSSE+KHDA	–	**98.1±0.5**	**94.3**

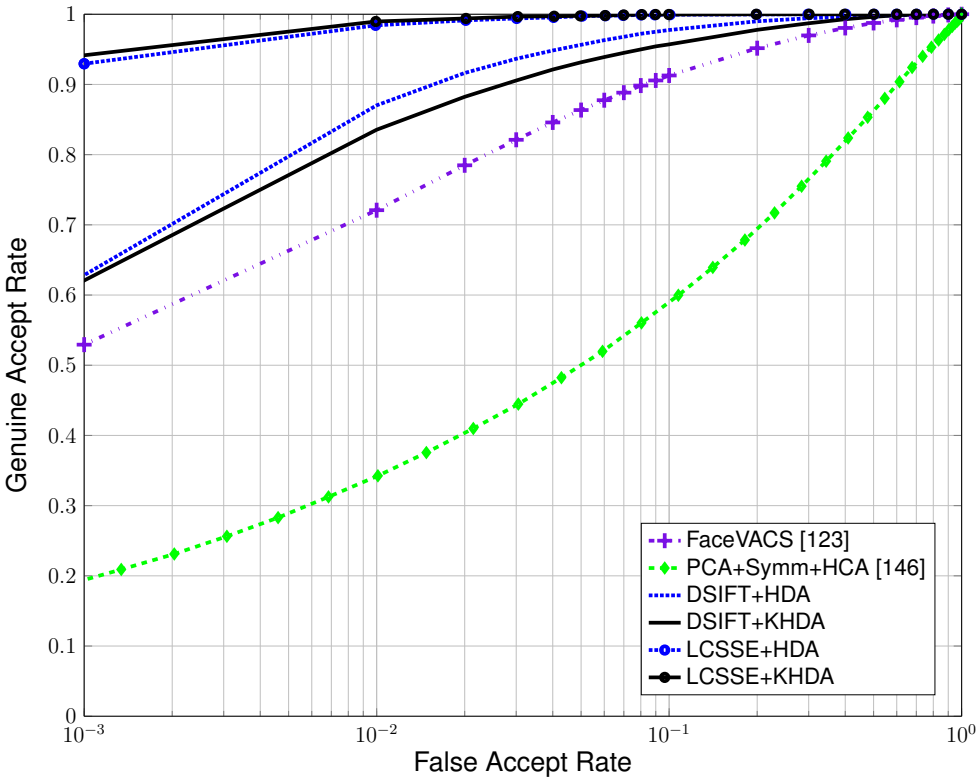


Figure 3-6: ROC curves on the CASIA NIR-VIS-2.0 database [146].

with the results reported in literature. The baseline results of (PCA+Symmetry+HCA) are provided with the dataset [146]. Along with that, we have evaluated the performance of one of the leading commercial off-the-shelf (COTS) face recognition system, FaceVACS[3]. Comparative analysis is shown with 20 recently published results for heterogeneous matching approaches, namely correlation component analysis (CCA [171]), subclass discriminant analysis (SDA [172]), partial least square (PLS [135]), couple-discriminative feature encoding (CDFE [132]), multi-view discriminant analysis (MvDA [140]), generalized multi view linear discriminant analysis (GMLDA [141]) and marginal fisher analysis (GMMFA [141]), coupled discriminant feature descriptor (C-DFD [173]), coupled discriminant feature learning (CDFL[121]), coupled compact binary face descriptor and its variants (C-CBFD, C-CBFD+LDA [147]), deep learning based shared representation [128], joint dictionary learning based approach [148], and deep learning based approaches [128]–[131], [174][4]. For experiments pertaining to KHDA, the Gaussian kernel function

[3]http://www.cognitec.com/technology.html

[4]Due to unavailability of the codes, comparison is shown with published results on official protocol.

Table 3.6: Rank-1 identification accuracy of the proposed HDA, KHDA and existing algorithms on CMU-MultiPIE database with different gallery and probe image sizes. Two top performing approaches are highlighted in each cross-resolution setting.

Resolution		**CTL**	**FaceVACS**	**DSIFT [168]**		**LCSSE [165]**	
Gallery	**Probe**	**[177], [178]**	**[177], [178]**	**HDA**	**KHDA**	**HDA**	**KHDA**
$\mathbf{216 \times 216}$	$\mathbf{72 \times 72}$	81.0	**99.5**	94.1	95.4	95.8	**97.0**
	$\mathbf{48 \times 48}$	79.7	**98.1**	92.4	94.1	93.7	**95.3**
	$\mathbf{32 \times 32}$	65.3	**97.4**	89.0	90.7	92.0	**93.2**
	$\mathbf{24 \times 24}$	37.7	54.5	87.3	85.7	**89.0**	**89.5**
	$\mathbf{16 \times 16}$	23.6	10.9	37.6	37.6	**61.2**	**62.5**
$\mathbf{72 \times 72}$	$\mathbf{48 \times 48}$	92.3	92.7	95.4	96.2	**96.6**	**97.0**
	$\mathbf{32 \times 32}$	84.1	84.3	92.4	**96.2**	92.8	**96.6**
	$\mathbf{24 \times 24}$	77.4	78.5	89.0	**91.6**	93.2	**94.1**
	$\mathbf{16 \times 16}$	72.4	72.8	44.3	54.9	**73.4**	**75.1**
$\mathbf{48 \times 48}$	$\mathbf{32 \times 32}$	61.8	96.8	95.4	**97.1**	96.2	**97.9**
	$\mathbf{24 \times 24}$	57.1	75.9	95.4	94.9	**96.6**	**97.5**
	$\mathbf{16 \times 16}$	32.9	6.4	73.8	71.3	**77.2**	**78.1**
$\mathbf{32 \times 32}$	$\mathbf{24 \times 24}$	45.7	78.4	94.9	94.5	**95.8**	**96.2**
	$\mathbf{16 \times 16}$	28.1	5.4	88.6	86.1	**90.3**	**91.1**
$\mathbf{24 \times 24}$	$\mathbf{16 \times 16}$	43.2	16.3	85.7	85.2	**87.3**	**89.0**

$k(x, y) = \exp\left(\frac{|x-y|^2}{2t}\right)$ is utilized and the parameter t is tuned from the development set.

Table 3.5 shows that with pixel values as input, the proposed HDA approach outperforms other existing algorithms. For example, MvDA with pixel values yields 41.6% rank-1 identification accuracy and 19.2% GAR at 0.1%FAR, whereas, the proposed approach yields similar rank-1 accuracy with lower standard deviation and much higher GAR of 31.4%. Further, Table 3.5 and Fig. 3-6 clearly demonstrate the performance improvement due to the proposed HDA and its non-linear kernel variant (KHDA). KHDA with learnt representation LCSSE outperforms all the existing algorithms in both identification and verification scenarios. It yields rank-1 identification accuracy of 98.1% (around 2.5% higher than the previous best reported result) and 94.3% GAR at 0.1% FAR. Using DSIFT features with the proposed KHDA also yields results comparable to other non-deep learning based approaches.

3.3.2 Cross Resolution Face Matching

Cross resolution face recognition entails matching high resolution gallery images with low resolution probe images. The cross resolution matching problem arises when the face images are either captured from different distances and/or camera sensor are of different resolutions. In this scenario, high resolution and low resolution are considered as two different *views* of a face image. For cross-resolution face matching, Bhatt *et al.* [177], [178] and Lie *et al.* [119], [137] have shown state-of-the-art results using co-transfer learning (CTL) and coupled discriminant analysis, respectively. The co-transfer learning algorithm [177], [178] combines co-training and transfer learning approaches for cross-resolution face matching. On the other hand, Lie *et al.* [119] presented locality constraint based coupled discriminant analysis with two variations: locality constraint in kernel space-coupled discriminant analysis (LCKS-CDA) and LCKS-coupled spectral regression (LCKS-CSR). Both the papers show results on the CMU Multi-PIE database [61] containing images pertaining to 337 individuals with pose, illumination, and expression variations. However, both the researchers have followed different protocols. In this research, we demonstrate the effectiveness of the proposed approach with the protocol followed by Bhatt *et al.* [177].

Experimental Protocol [177], [178]: In many face recognition applications, it is generally assumed that the gallery contains high resolution images. Therefore, combinations of gallery and probe set pairs are created such that probe images have lower resolution than gallery images. Each image is resized to six different resolutions: 16×16, 24×24, 32×32, 48×48, 72×72, and 216×216. In total, 15 cross-matching scenarios are considered. For example, with 216×216 size gallery set, five experiments pertaining to 72×72, 48×48, 32×32, 24×24, and 16×16 size probe sets are performed. For every person, two images are selected and images pertaining to 100 subjects are utilized for training, whereas the remaining 237 subjects are utilized for testing. The results are reported in Table 3.6 and Fig. 3-7. Since the protocol [177], [178] does not involve cross-validation, error intervals are not reported.

It can be seen that LCSSE+KHDA outperforms the co-transfer learning [177], [178] in all the cross-resolution matching scenarios. For example, when 48×48 pixels gallery images are matched with probe images of 32×32, 24×24, and 16×16 pixels, performance improvement of about

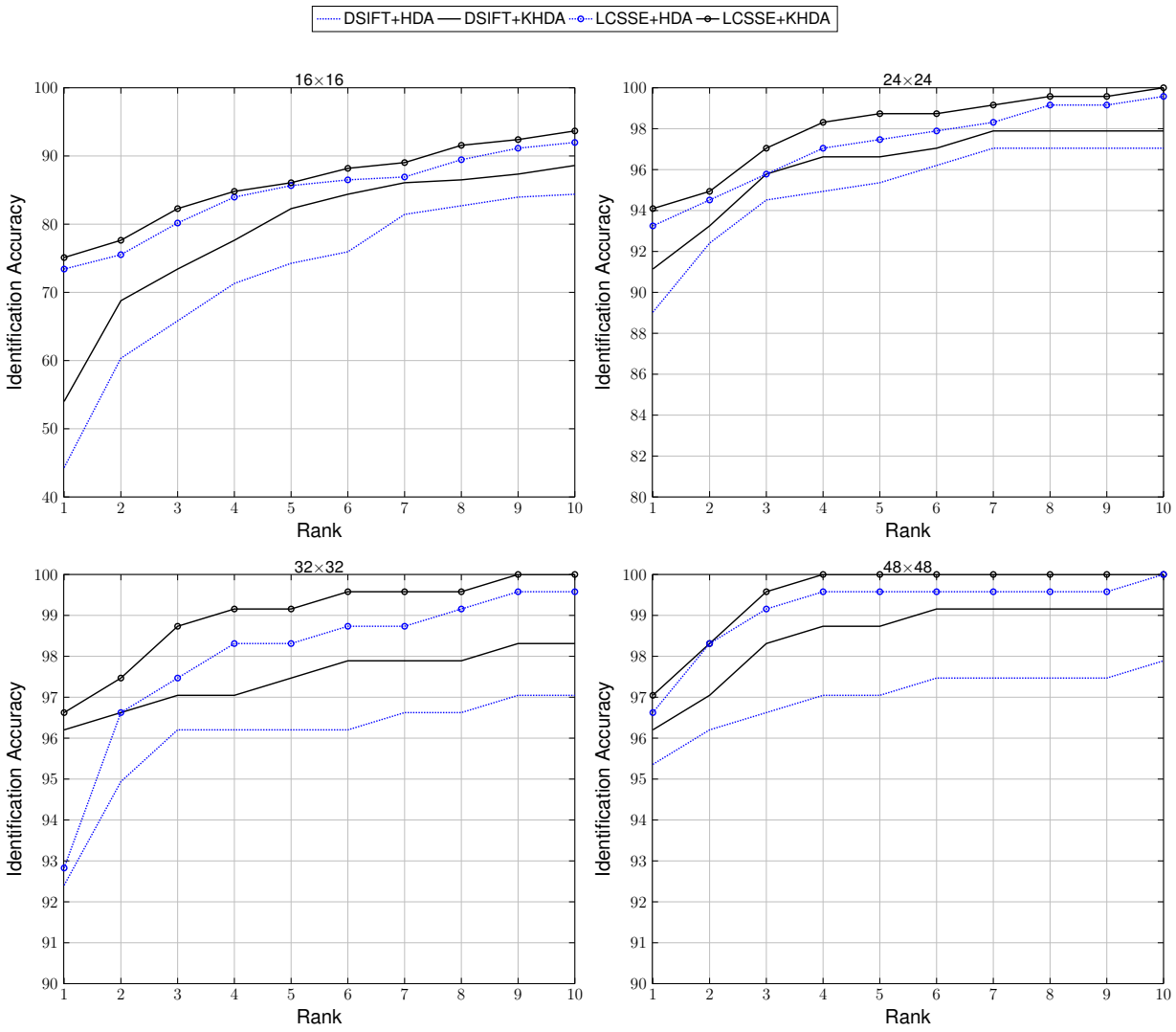


Figure 3-7: CMC curves for cross-resolution matching (Probe: 48×48, 32×32, 24×24, 16×16, Gallery: 72×72) on the CMU-MultiPIE database [61].

30%-40% is observed. Analyzing the results across resolutions show that the accuracy reduces with increase in resolution difference between the gallery and probe images. When compared with LCSSE alone (without HDA/KHDA), as shown in Fig. 3-8, we observe that KDHA improves the performance of LCSSE by 2.8 - 8.9%.

FaceVACS yields impressive performance when the size of both gallery and probe are higher than 32×32. However, the performance deteriorates significantly with decrease in the gallery image size and with increase in the resolution difference. Generally, the performance of the proposed HDA and/or KHDA is less affected due to resolution difference in comparison to FaceVACS,

Figure 3-8: Performance improvement due to HDA and KHDA with LCSSE features on the CMU Multi-PIE database. Size of gallery images is 48×48, whereas probe sizes are 32×32, 24×24, 16×16.

and CTL. We have also observed that for cross-resolution face recognition, LCSSE shows higher accuracies compared to DSIFT with a difference of up to 25%.

3.3.3 Digital Photo to Composite Sketch Face matching

In many law enforcement and forensic applications, software tools are used to generate composite sketches based on eye-witness description and the composite sketch is matched against a gallery of digital (mugshot) photographs. While there is some research on forensic hand-drawn sketches [179], [180], the research pertaining to composite sketch matching is relatively less explored. Han *et al.* [161] presented a component based approach followed by score fusion for composite to photo matching. Later, Mittal *et al.* [162], [181]–[183] and Chugh *et al.* [184] presented learning based algorithms for the same. Klum *et al.* [185] presented FaceSketchID for matching composite sketches to photos.

For this set of experiments, we perform experiments on the e-PRIP composite sketch dataset [161], [162]. The dataset contains composite sketches of 123 face images from the AR face dataset [40]. It contains the composite sketches created using two softwares, Faces and IdentiKit[5]. The PRIP dataset [161] originally has composite sketches prepared by a Caucasian user (with IdentiKit

[5]Faces: http://www.iqbiometrix.com, IdentiKit: http://www.identikit.net

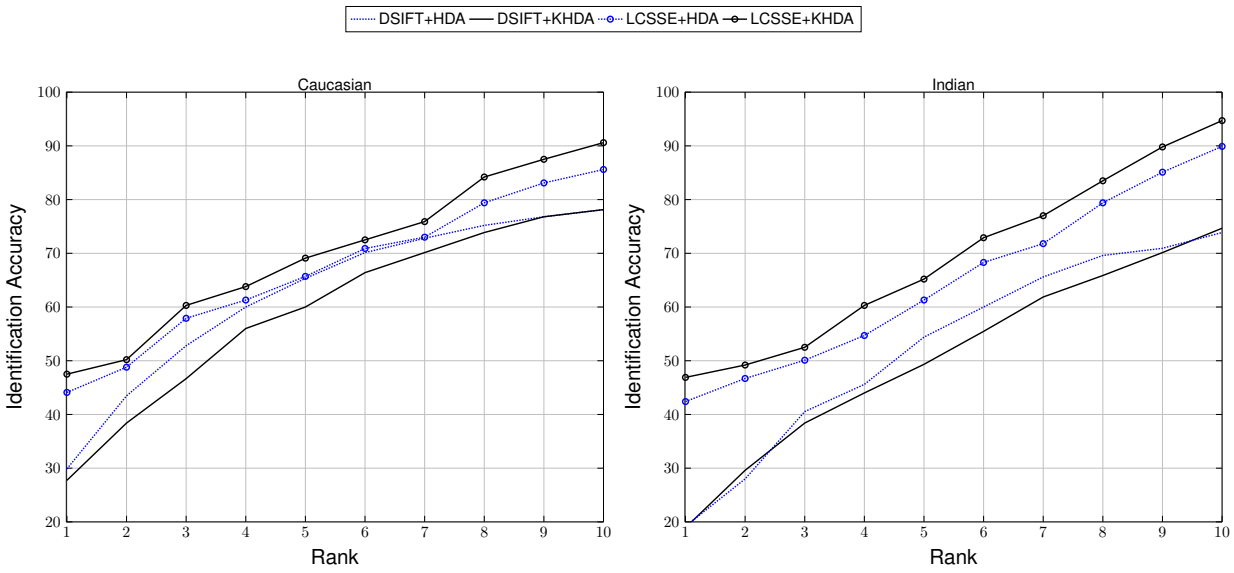


Figure 3-9: CMC curve for composite sketch to digital photo matching on the e-PRIP composite sketch dataset [161], [162].

and Faces softwares) and an Asian user (with Faces software). Later, the dataset is extended by Mittal *et al.* [162] by adding composite sketches prepared by an Indian user (with Faces software) which is termed as the e-PRIP composite sketch dataset. In literature, it has been shown that composite sketches prepared by Caucasian and Indian users using Faces software yield better results compared to other sets [162], [181]. Therefore, in this work, we use composite sketches prepared using Faces software by the Caucasian and Indian users. The experiments are performed with the same protocol as presented by Mittal *et al.* [162]. The dataset is divided into 40% training (48 subjects) and 60% testing (75 subjects), with random sampling based five times cross validation. Average identification accuracies at Rank-10 are reported in Table 3.7 and Fig. 3-9 shows the corresponding CMC curves.

With the above mentioned experimental protocol, one of the best results in literature have been reported by Mittal *et al.* [183] with rank-10 identification accuracies of 59.3% (Caucasian) and 58.4% (Indian). Saxena and Verbeek [174] have shown results with Indian users only and have achieved 65.5% rank-10 accuracy. As shown in Table 3.7, the proposed approaches, HDA and KHDA, with both DSIFT and LCSSE improve the performance significantly. Compared to existing algorithms, DSIFT demonstrates an improvement in the range of 11–23% while LCSSE+HDA and LCSSE+KHDA improve the rank-10 accuracy by approximately 30% with respect to state-

Table 3.7: Rank-10 Identification accuracy for composite sketch to photo matching. The results marked with $*$ are reported by Mittal *et al.* [162].

Algorithm	**Rank-10 Accuracy (%)**	
	Faces (Caucasian)	**Faces (Indian)**
Mittal *et al.* [181]*	32.4±2.4	30.3±1.7
Mittal *et al.* [162]*	51.9±1.2	53.3±1.4
Mittal *et al.* [182]	56.0±2.1	60.2±2.9
Mittal *et al.* [183]	59.3±0.8	58.4±1.1
COTS [162]*	11.3±2.1	9.1±1.9
Saxena and Verbeek [174]	-	65.6±3.7
DSIFT only	67.5±5.8	51.7±4.0
DSIFT+HDA	79.5±2.8	73.9±5.8
DSIFT+KHDA	78.6±3.4	74.6±3.8
LCSSE only	**68.0±2.6**	**65.3±4.1**
LCSSE+HDA	**85.6±1.3**	**89.0±1.5**
LCSSE+KHDA	**89.6±1.9**	**94.7±1.0**

of-the-art. Similar to previous results, this experiment also shows that application of HDA/KHDA improves the results of DSIFT and LCSSE.

3.4 Comparison with Related Approaches

As discussed in literature review, several discriminant analysis based approaches have been proposed. Here, we compare and contrast the similarities and differences, with selected approaches, in terms of their objective functions and experimental results.

HDA vs SDA: Within the framework of Subclass discriminant analysis [172], samples of each class with two views, may be viewed as each class having two subclasses; with each subclass representing one view. With this perspective the between-class scatter matrix of HDA can be closely related to its definition in subclass discriminant analysis. However, the computation of total-scatter matrices are different. Moreover, SDA is not designed specifically for heterogeneous recognition tasks. Empirical comparison is shown in Table 3.5. For VIS-NIR matching, SDA yields 75.7% rank-1 accuracy and 54.8% GAR@0.1% FAR, whereas, with same features, HDA and KHDA yield 81.0%, 83.1% rank-1 identification accuracy, and 62.8%, 62.1% GAR@0.1% FAR, respec-

tively.

HDA vs CDA: Coupled discriminant analysis [119], [137] utilizes inter-view and intra-view pairs of samples (or class means). In other words, the S_b scatter matrix constitutes of inter-view variation (S_b^{gp} and S_b^{pg}) and intra-view variation (S_b^{pp} and S_b^{gg}) components; similarly for within-class scatter matrix S_w. In a broader perspective, HDA can be seen as rejecting the intra-view information and preserving only the inter-view information. Empirically (not reported here) we observe that compared to both the variants of CDA, the proposed HDA and KHDA yield approximately 2%-15% improvement in rank-1 identification accuracy for the most of the cross-resolution face matching scenarios.

HDA vs MvDA: Multi-view discriminant analysis [140] defines the within-class scatter matrix as $\sum_{i=1}^{c}\sum_{j=1}^{v}\sum_{k=1}^{n_{ij}}(y_{ijk}-\mu_i)(y_{ijk}-\mu_i)^T$ and between-class scatter matrix as $\sum_{i=1}^{c}(\mu_i-\mu)(\mu_i-\mu)^T$; where y_{ijk} is the projection of k^{th} sample of j^{th} view of i^{th} class. Thus, it aims at obtaining view specific projection directions such that (i) the samples of a class are closer to its class mean (across views: μ_i), and (ii) class means (across views: μ_i) are away from the overall mean (μ) in the projected space. Both the objectives are different from HDA where (a) the samples of one view of a class are brought closer to mean of another view of the same class (μ_i^a, μ_i^b), and (b) mean of one view of a class (μ_i^a) is pulled apart from means of other view of other classes ($\mu_l^b, i \neq l$). The comparison of MvDA with the proposed HDA and KHDA in terms of results is shown in Table 3.5. The results of HDA and KHDA are significantly better than MvDA.

HDA vs GMA [141]: Generalized multiview analysis [141] optimizes the $w_1^T A_1 w_1 + \beta w_2^T A_2 w_2 + 2\alpha w_1^T Z_1 Z_2^T w_2$ under the constraint of $w_1^T B_1 w_1 + \gamma w_2^T B_2 w_2 = 1$, where A_i and B_i represent within-view inter- and intra- class information, and Z_1 and Z_2 consist of samples of view 1 and 2, respectively. The objective function encodes difference of intra-view samples and correlation of inter-view samples. The last term in the optimization function tries to maximize covariance between the samples from different views to obtain directions to achieve closeness between multi-view samples of the same class. In other words, GMA can be considered as optimizing the between and within-class scatter of individual views and the cross-view correlation in a weighted manner.

In case of HDA, the objective function is catered towards obtaining discriminative projections for inter-view scenarios. Table 3.5 shows comparative results of both the GMA based approaches, GMLDA and GMMFA, for VIS-NIR matching. GMLDA and GMMFA [141] achieve 23.7% and 23.8% rank-1 accuracy [121], [147], which is significantly lower compared to our results.

3.5 Summary

In this research, we have proposed a discriminant analysis approach for heterogeneous face recognition. We formulate heterogeneous discriminant analysis which encodes view labels and has the objective function optimized for heterogeneous matching. Based on the analytical solution, we propose its kernel extension, KHDA. Experimental results are performed on three heterogeneous face matching problems, namely, visible to NIR matching, cross resolution matchings, and digital photo to sketch, with hand-crafted DSIFT and deep learning based LCSSE. The results consistently show that the proposed modification in the classical discriminant analysis technique exhibits significantly improved recognition performance. For all three case studies, the proposed HDA and KHDA are among the top performing approaches.

Chapter 4

Incremental Semi-supervised Discriminant Analysis for Face Recognition

Discriminant analysis (DA) [186] based classifiers have found their utility in wide range of problems such as image retrieval [187] and face recognition [16]. Linear discriminant analysis (LDA) [188] and its variants have been efficiently used in various pattern classification problems [172], [189]–[191]. Some of the most interesting successors of LDA are kernel LDA [192], maximum margin criterion based discriminant function [193], and graph-embedding [190], [194] based algorithms.

The formulations of DA techniques, typically, require labeled training data. In certain applications, such as image retrieval and object classification, it is difficult to obtain large labeled data. However, large amount of unlabeled data is easily available. To address this aspect, researchers introduced semi-supervised learning in discriminant analysis [195]–[198]. The paradigm utilizes labeled as well as large amount of unlabeled training data to learn the model [199]. Semi-supervised learning is very important in addressing the labeled data related limitation, as it learns a model from labeled as well as large amount of unlabeled training data. Therefore, semi-supervised learning approaches have been proposed in discriminant analysis. Generally, existing semi-supervised incremental learning algorithms first learn the model using labeled data, which is followed by classification of unlabeled data [197], [198], [200]. Either a new classification model is learned or existing model is updated using the confidently classified unlabeled data samples. Therefore, these set of algorithms create pseudo labeled data from unlabeled data, and use an existing supervised

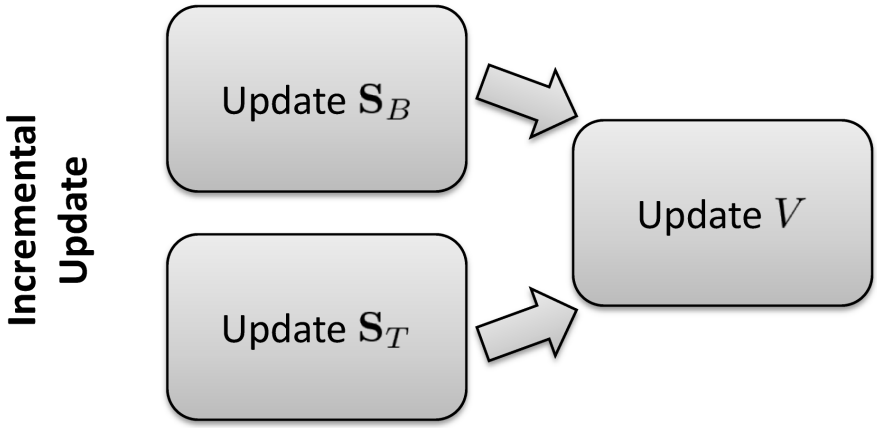


Figure 4-1: Traditional incremental discriminant approaches, such as Kim *et al.* [207], [208] and Lamba *et al.* [211], update between-class and overall variability. New eigenmodels of $\mathbf{S}_B$ and $\mathbf{S}_T$ are learned from incremental batch, which are merged with corresponding existing eigenmodels. Discriminating components $\mathbf{V}$ are obtained from merged eigenmodels of $\mathbf{S}_B$ and $\mathbf{S}_T$.

learning framework. However, this approach requires to iteratively learn the model which might be time consuming. Cai *et al.* [195] proposed semi-supervised discriminant analysis (SSDA) by utilizing unlabeled data for learning the regularized total scatter matrix. The regularization is performed using graph Laplacian of unlabeled training set which encodes the manifold assumption. Few other related semi-supervised learning algorithms are summarized in Table 4.1. Semi-supervised learning can be utilized in various research areas ranging from bioinformatics, speech recognition, natural language parsing, and spam filtering [199], [201].

Both the types of discriminant analysis algorithms, supervised (e.g. LDA) or semi-supervised (e.g. SSDA), are usually trained in batch mode. In many real world applications, it is likely that whole labeled training set is not available before hand; rather the training data is obtained incrementally. The batch learning algorithms have a major limitation related to very limited provision for updating discriminant components by incorporating the newly available training samples only. To obtain a new model, the discriminant classifier has to be learned from the merged data i.e. both original and incremental training data. Since the core of every discriminant analysis objective function contains an eigenvalue decomposition problem, learning a new classifier from merged data has cubic time complexity. Further, as SSDA encodes the data in the form of graphs, graph adjacency matrix has to be obtained from the merged data to update the model. This presents an additional challenge for learning a new model. The challenge can be addressed using incremental

Table 4.1: Literature review of the related research pertaining to incremental learning and semi-supervised learning algorithms related to discriminant analysis.

Algorithm	**Year**	**Description**
Semi-supervised Leaning		
SSDA [195]	2007	Semi-supervised discriminant analysis based on LDA which uses unlabeled set for estimating total scatter
SELF [196]	2010	A semi-supervised extension of local fisher discriminant analysis that preserves the global structure of the unlabeled data.
SSGDA [197]	2011	Confidently classified unlabeled data samples are utilized with pseudo labels in generalized discriminant analysis.
Byun [198]	2012	Utilizes pseudo labels of only those unlabeled samples that are expected to reduce errors.
Incremental Leaning		
IPCA [202]	2000	Proposed algorithm for merging eigenpaces of total scatter matrices
ILDA-Pang[203]	2005	Incrementally updates between- and within-class scatter.
GSVD-ILDA [204]	2008	Incremental version of LDA/GSVD [205]
LS-ILDA [206]	2009	Formulates ILDA in terms of least square solution by incrementally updating total scatter of mean centered data matrix.
IDCC [200]	2010	Incremental discriminant canonical correlation analysis by adapting the sufficient spanning set based merging of eigenspace [202]
ILDA [207], [208]	2007, 2011	Merging eigenspaces of between- and within-class scatter of existing and new batch for updating model.
I-CLDA[209], [210]	2012	Incremental complete linear discriminant analysis utilizing QR decomposition to obtain orthonormal projection directions.
ISDA (subclass)[211]	2012	Extension of ILDA[208] to incremental subclass discriminant analysis
ILDA-KT [212]	2012	Addressing concept drift in incremental learning using knowledge transfer
LS-LDA-CD [213]	2013	Addresses concept drift issue in least square LDA [206]
Chunk-IDR/QR [214]	2015	A time-effecient version of IDR-QR [215]
ILDA/QR [216]	2015	Utilization of QR decomposition of data matrix for incremental learning.
Proposed ISSDA	-	Extension of ILDA to semi-supervised discriminant analysis with reduced time complexity.

learning [206]–[208] where new training samples are incrementally incorporated into the classification model. The motivation of incremental learning for discriminant analysis is to be able to *update* the existing model using the newly available training samples with significantly less time complexity. Some existing contributions pertaining to incremental learning are summarized in Table 4.1. Kim *et al.* [207], [208] and Lamba *et al.* [211] utilized the eigenspace merging algorithm [202] to formulate incremental linear discriminant analysis (ILDA) and incremental subclass discriminant analysis, respectively. As shown in Figure 4-1, both the algorithms use the new training samples to update the between-class scatter matrix and the total scatter matrix individually, and learn the discriminating components . Liu *et al.* [206] proposed an incremental learning algorithm based on the least square formulation of LDA. Time complexity of the algorithm proposed in [206] is less than that of [207], [208], however the space complexity of the former is more as it requires to store the entire data matrix or total scatter matrix as part of the classification model.

To mitigate the above mentioned challenges, this research presents an incremental semi-supervised discriminant analysis (ISSDA) algorithm. We address the problem with two reasonable assumptions: large unlabeled training data is available offline and labeled data is received incrementally. The proposed algorithm aims at reducing the computational complexity of the incremental update process by utilizing the unlabeled dataset for robust data statistics estimation. The major contributions of this research are:

- showcasing that large unlabeled training set can be leveraged to efficiently estimate the total scatter matrix,
- utilization of manifold regularization of robust estimation of total scatter matrix, and
- sufficient spanning set representation [207], [208] based incremental learning approach which requires to update only the between-class scatter matrix and not the total scatter matrix.

The effectiveness of the proposed algorithm is evaluated for face recognition application. The performance is evaluated by comparing the accuracy, time and consistency of the proposed incremental algorithm with the corresponding batch learning model. Evaluations to understand the effects of the manifold regularizer and unlabeled data size are also performed. Further, the effect of updating the model with incremental batch consisting of samples of new classes is also studied.

4.1 Incremental Semi-supervised Discriminant Analysis

Discriminant analysis based approaches have a fundamental objective of maximizing the inter-class variation and minimizing the intra-class variation. In case of linear discriminant analysis, inter-class variability is modeled in terms of between-class scatter matrix $\mathbf{S}_B$ and intra-class variability is modeled in terms of within-class scatter matrix $\mathbf{S}_W$ [188],

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \text{ and} \tag{4.1}$$

$$\mathbf{S}_W = \sum_{i=1}^{c} \sum_{x_j \in \mathbf{X}_i} (x_i - \boldsymbol{\mu}_i)(x_i - \boldsymbol{\mu}_i)^T \tag{4.2}$$

Here, c is the number of classes, n is the total number of samples, n_i is the number of samples in i^{th} class, $\mathbf{X} = [\mathbf{X}_1 \dots \mathbf{X}_c]$ is the data matrix, $\mathbf{X}_i$ is the set of samples belonging to i^{th} class, $\boldsymbol{\mu}_i (= \frac{1}{n_i} \sum_{x_j \in \mathbf{X}_i} x_j)$ is the mean of i^{th} class, and $\boldsymbol{\mu} (= \frac{1}{n} \sum_{i=1}^{c} \sum_{x_j \in \mathbf{X_i}} x_j)$ is the mean of all the data samples. The objective is to find the set of projection directions $\mathbf{V}$ such that,

$$\mathbf{V} = \arg\max \frac{|\mathbf{V}^T \mathbf{S}_B \mathbf{V}|}{|\mathbf{V}^T \mathbf{S}_W \mathbf{V}|} \tag{4.3}$$

In many applications, $\mathbf{S}_W$ can be singular; therefore, the following equivalent criterion [188] can be used:

$$\mathbf{V} = \arg\max \frac{|\mathbf{V}^T \mathbf{S}_B \mathbf{V}|}{|\mathbf{V}^T (\mathbf{S}_W + \mathbf{S}_B) \mathbf{V}|} = \arg\max \frac{|\mathbf{V}^T \mathbf{S}_B \mathbf{V}|}{|\mathbf{V}^T \mathbf{S}_T \mathbf{V}|} \tag{4.4}$$

Where, $\mathbf{S}_T$ is the total scatter matrix defined as,

$$\mathbf{S}_T = \sum_{i=1}^{c} \sum_{x_j \in \mathbf{X}_i} (x_i - \boldsymbol{\mu})(x_i - \boldsymbol{\mu})^T \tag{4.5}$$

Also note that, Eq. 4.5 suggests that $\mathbf{S}_T$ depends on the global mean; not on the class means, and can be computed without having the knowledge of class labels. The criterion, Eq. 4.4, can be modeled into a generalized eigenvalue problem as

$$\mathbf{S}_B \mathbf{V} = \lambda \mathbf{S}_T \mathbf{V} \tag{4.6}$$

where, the solution $\mathbf{V}$ contains the most discriminant projection directions.

4.1.1 Semi-Supervised Discriminant Analysis

Since the global statistics may be estimated with limited training data, the Fisher criterion in Eq. 4.4 is prone to over fitting. Therefore, the following regularization criterion is often used [217].

$$\mathbf{V} = \arg\max \frac{|\mathbf{V}^T \mathbf{S}_B \mathbf{V}|}{|\mathbf{V}^T \mathbf{S}_T \mathbf{V} + \beta_1 \mathbf{R}(\mathbf{V})|} \tag{4.7}$$

where, $\mathbf{R}(\mathbf{V})$ is the regularizer function and β_1 controls the weight given to it. Cai *et al.* [195] proposed to use a graph embedding based regularizer

$$\mathbf{R}(\mathbf{V}) = \mathbf{V}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{V} \tag{4.8}$$

where, $\mathbf{L}$ is the Laplacian of the graph built by considering each sample x_i as the node and the edges describe the *connectivity* of the two samples. The edges are encoded using an adjacency matrix $\mathbf{W}$ such that the entry $\mathbf{W}_{ij}$ is $e^{\gamma||x_i - x_j||_2^2}$ if x_j is the neighborhood of x_i or vice-versa and in other cases $\mathbf{W}_{ij}$ is zero. The graph Laplacian $\mathbf{L}$ can be computed as $\mathbf{L} = \mathbf{W} - \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix of row sum (or column sum) of $\mathbf{W}$ $\left(\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}\right)$ [218], [219].

Manifold assumption [218]–[220] expects data samples to be on a surface, whereas data distribution assumption [221] expects data samples to follow certain distributions. Since manifold assumption is rather relaxed and more effective in data representation, it is utilized for regularization in SSDA. Similarly, [196], [222]–[226] have utilized the manifold assumption to better model the classifier.

Substituting Eq. 4.8 in Eq. 4.7 results in

$$\mathbf{V} = \arg\max \frac{|\mathbf{V}^T \mathbf{S}_B \mathbf{V}|}{|\mathbf{V}^T (\mathbf{S}_T + \beta_1 \mathbf{X} \mathbf{L} \mathbf{X}^T) \mathbf{V}|} \tag{4.9}$$

However, the denominator in this modified criterion is not guaranteed to be nonsingular. Therefore,

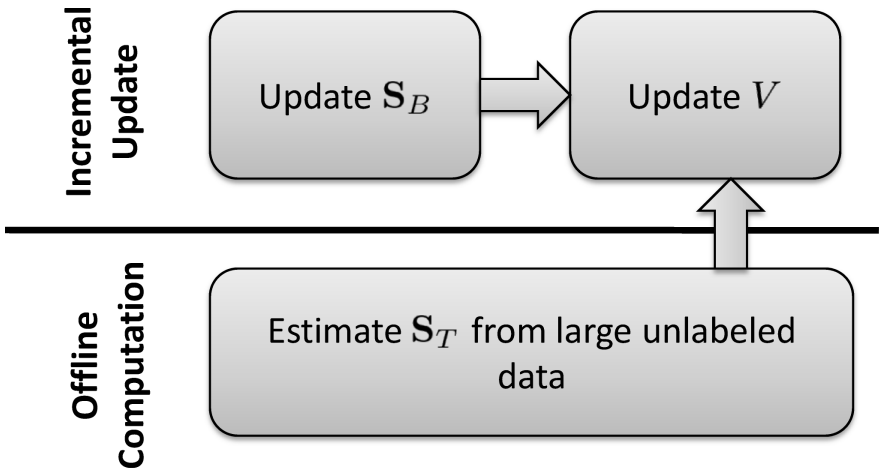


Figure 4-2: The proposed approach incrementally learns the between-class variability and uses unlabeled data to learn the overall variability. Eigenmodel of $\mathbf{S}_B$ is learned from incremental batch and merged with the existing eigenmodel. New discriminating components $\mathbf{V}$ are obtained using updated eigenmodel of $\mathbf{S}_B$ and offline estimated eigenmodel of $\mathbf{S}_T$.

a small positive value $\beta_2 > 0$ is added to the diagonal elements making the criterion

$$\mathbf{V} = \arg\max \frac{|\mathbf{V}^T \mathbf{S}_B \mathbf{V}|}{|\mathbf{V}^T (\mathbf{S}_T + \beta_1 \mathbf{X}\mathbf{L}\mathbf{X}^T + \beta_2 \mathbf{I})\mathbf{V}|} \tag{4.10}$$

Similar to Eq. 4.6, Eq. 4.10 results in the eigenvalue decomposition problem as follows

$$\mathbf{S}_B \mathbf{V} = \lambda \left(\mathbf{S}_T + \beta_1 \mathbf{X}\mathbf{L}\mathbf{X}^T + \beta_2 \mathbf{I}\right) \mathbf{V} \tag{4.11}$$

The denominator of Eq. 4.10 encodes the overall/total variability irrespective of the class labels, in other words “the expected variation among samples". This reflects in the fact that no term in the denominator depends on class specific statistics. Interestingly, this overall variability can be learned from an unlabeled dataset [195]. Let $\mathbf{X}^{(u)}$ be the unlabeled data matrix, $\mathbf{S}_T^{(u)}$ be its total scatter matrix, and $\mathbf{L}^{(u)}$ be the corresponding graph Laplacian. Eq. 4.11 can be written as,

$$\mathbf{S}_B \mathbf{V} = \lambda \left(\mathbf{S}_T^{(u)} + \beta_1 \mathbf{X}^{(u)} \mathbf{L}^{(u)} \mathbf{X}^{(u)^T} + \beta_2 \mathbf{I}\right) \mathbf{V} \tag{4.12}$$

4.1.2 Incremental Learning

To incorporate incremental learning in Eq. 4.12, as shown in Figure 4-1 the between-class scatter and the regularized total scatter matrices needs to be updated, and accordingly new discriminating projection directions are to be computed. This update of $\mathbf{S}_B$ and $\mathbf{S}_T$ should reflect the changes in the class specific statistics $\boldsymbol{\mu}_i$ and the global statistic $\boldsymbol{\mu}$, respectively. While a model can be learned by recomputing $\mathbf{S}_B$ and $\mathbf{S}_T$ along with solving the eigenvalue decomposition in Eq. 4.6 from the combined training set, this approach can be very expensive with large data. Therefore, incrementally updating scatter matrices is preferable than retraining the model with cumulated data.

As illustrated in Figure 4-2, we propose to estimate the total scatter from large unlabeled data. It is our assertion that the size of unlabeled dataset affects total variability estimation. With bigger unlabeled dataset, $\mathbf{S}_T$ can more accurately estimate the total scatter of the population and further addition of new samples should not change it significantly. Thus, *if $\mathbf{S}_T$ is learned offline from sufficiently large data, it is not required to be updated with incremental learning.* Therefore, the estimate of total scatter can be precomputed and only $\mathbf{S}_B$ is to be updated incrementally. We utilize sufficient spanning set representation of scatter matrices [207], [208], to obtain incrementally updatable model. This section explains the proposed approach of incrementally updating the semi-supervised learning based model of ISSDA.

Sufficient Spanning Set Representations and Model Update

Let the existing model be trained on M_1 samples and the incremental batch contains M_2 new samples. Let $\mathbf{S}_{B,1}$ and $\mathbf{S}_{B,2}$ be the between-class scatter matrices of the existing batch and the incremental batch respectively. The eigenspace model of the between class scatter matrix may be represented as $\{\boldsymbol{\mu}_i, M_i, \mathbf{Q}_i, \boldsymbol{\Delta}_i, n_i, \boldsymbol{\alpha}_i\}_{i=1,2}$, where i represents the i^{th} incremental batch of new data points, and for corresponding batches $\boldsymbol{\mu}$, M, $\mathbf{Q}$, and $\boldsymbol{\Delta}$ are the mean, number of samples, eigenvector matrix, and eigenvalue matrix respectively. $\boldsymbol{\alpha}_i$ is the vector of coefficients to represent the mean of the j^{th} class of i^{th} batch as $\mathbf{m}_{ij} \simeq \boldsymbol{\mu}_i + \mathbf{Q}_i \alpha_{ij}$. Note that the original between-class scatter matrix can be approximated by using the corresponding eigenmodel as $\mathbf{S}_{B,i} = \mathbf{Q}_i \boldsymbol{\Delta}_i \mathbf{Q}_i^T$. The eigenvector matrix $\mathbf{Q}_i$ does not contain *all* the eigenvectors of the matrix, rather it contains the

set of eigenvectors which are *sufficient* enough to reconstruct the matrix [207], [208]. In this way, eigenmodel inherently encompasses the idea of sufficient spanning set, i.e. the set of sufficient eigenvectors that span the space of scatter matrix.

Let $\mathbf{S}_{B,3}$ be the between-class scatter matrix of the merged training sets of both the batches,

$$\mathbf{S}_{B,3} = \mathbf{S}_{B,1} + \mathbf{S}_{B,2} + \frac{M_1 M_2}{M_1 + M_2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T + \mathbf{A} \tag{4.13}$$

where,

$$\mathbf{A} = \sum_{k \in S} \frac{-n_{1k} n_{2k}}{n_{1k} + n_{2k}}(\mathbf{m}_{2k} - \mathbf{m}_{1k})(\mathbf{m}_{2k} - \mathbf{m}_{1k})^T, \tag{4.14}$$

S is the set of common classes between the existing batch and new batch, and n_{ij} is the number of samples in the j^{th} class in i^{th} batch. The solution to finding the merged eigenmodel follows three steps:

1. The orthogonalization of $\mathbf{Q}_1, \mathbf{Q}_2$, and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ gives the set of orthonormal basis $\boldsymbol{\Psi}$, since the principal components ($\mathbf{Q}_1$ and $\mathbf{Q}_2$) and mean difference vector ($\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$) can span the data space of between-class scatter matrix [207], [208]. Moreover, $\mathbf{Q_3} = \boldsymbol{\Psi}\mathbf{R}$, where $\mathbf{R}$ is the rotation matrix. Orthogonalization can be performed using various methods, one such method is Gram-Schmidt orthonormalization [227], which computes the matrix $\boldsymbol{\Psi}$ in the first step.

2. Since $\mathbf{Q}_3$ is the eigenvector matrix of $\mathbf{S}_{B,3}$,

$$\begin{aligned} \mathbf{S}_{B,3} &= \mathbf{Q}_3 \boldsymbol{\Delta}_3 \mathbf{Q}_3^T && (4.15) \\ &= \boldsymbol{\Psi}_3 \mathbf{R} \boldsymbol{\Delta}_3 \mathbf{R}^T \boldsymbol{\Psi}_3^T && (4.16) \\ \implies \boldsymbol{\Psi}_3^T \mathbf{S}_{B,3} \boldsymbol{\Psi}_3 &= \mathbf{R} \boldsymbol{\Delta}_3 \mathbf{R}^T && (4.17) \end{aligned}$$

   where, the term on the left can be approximated without computing the matrix $\mathbf{S}_{B,3}$ [202]. Eigenvalue decomposition of this approximation yields the rotation matrix $\mathbf{R}$ and diagonal matrix containing eigenvalues $\boldsymbol{\Delta}_3$.

3. The eigenvector is constructed as $\mathbf{Q}_3 = \boldsymbol{\Psi}_3 \mathbf{R}$. The remaining parts of the eigenmodel are

computed as

$$M_3 = M_1 + M_2 \tag{4.18}$$

$$n_{3j} = n_{1j} + n_{2j} \tag{4.19}$$

$$\boldsymbol{\mu}_3 = \frac{M_1\boldsymbol{\mu}_1 + M_2\boldsymbol{\mu}_2}{M_3} \tag{4.20}$$

$$\boldsymbol{\alpha_{3j}} = \mathbf{Q}_3^T(\mathbf{m}_{3j} - \boldsymbol{\mu}_3) \tag{4.21}$$

$$\text{where, } \mathbf{m}_{3j} = \frac{n_{1j}\mathbf{m}_{1j} + n_{2j}\mathbf{m}_{2j}}{n_{3j}} \tag{4.22}$$

Similar to between-class scatter matrix, the regularized total-scatter matrix should also be represented using sufficient spanning sets. Since it is computed only once and not updated incrementally, it is sufficient to store only the eigenvectors and eigenvalues in its eigenmodel. The regularized total scatter matrix $\mathbf{S}_T^{(u)} + \beta_1\mathbf{X}^{(u)}\mathbf{L}^{(u)}\mathbf{X}^{(u)^T} + \beta_2\mathbf{I}$ may be represented using $\{\mathbf{P}, \mathbf{\Lambda}\}$, where $\mathbf{P}$ is the matrix containing leading eigenvectors of the scatter matrix and $\mathbf{\Lambda}$ is the diagonal matrix containing the corresponding eigenvalues. Note that the sufficiency of eigenvectors translates to the fact that

$$\mathbf{S}_T^{(u)} + \beta_1\mathbf{X}^{(u)}\mathbf{L}\mathbf{X}^{(u)^T} + \beta_2\mathbf{I} \simeq \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T \tag{4.23}$$

Finding Updated Discriminant Components

Having updated the eigenmodel of $\mathbf{S}_{B,3}$ incrementally, the next step is to find the updated discriminant components using the offline computed model $\{\mathbf{P}, \mathbf{\Lambda}\}$ representing the regularized total scatter and the updated model $\{\boldsymbol{\mu}_3, M_3, \mathbf{Q}_3, \mathbf{\Delta}_3, n_3, \boldsymbol{\alpha}_3\}$. Similar to Kim *et al.* [207], [208], discriminant components are computed using Eq. 4.24

$$\mathbf{V} = \mathbf{Z}\mathbf{\Omega}\mathbf{R} \tag{4.24}$$

where $\mathbf{Z} = \mathbf{P}\mathbf{\Lambda}^{-\frac{1}{2}}$, $\mathbf{\Omega}$ is the matrix containing basis vectors computed by orthogonalization of $\mathbf{Z}^T\mathbf{Q}_3$, and $\mathbf{R}$ is the matrix containing the eigenvectors of $\mathbf{\Omega}^T\mathbf{Z}^T\mathbf{Q}_3\mathbf{\Delta}_3\mathbf{Q}_3^T\mathbf{Z}\mathbf{\Omega}$ which is an approximation of $\mathbf{\Omega}^T\mathbf{Z}^T\mathbf{S}_{B,3}\mathbf{Z}\mathbf{\Omega}$. The initial batch and the incremental learning procedures of the proposed semi-supervised incremental discriminant analysis are summarized in Psuedocode 1. In the initial training a model $\mathbf{\Upsilon}_1$ is learned using labeled and unlabeled samples. The incremental

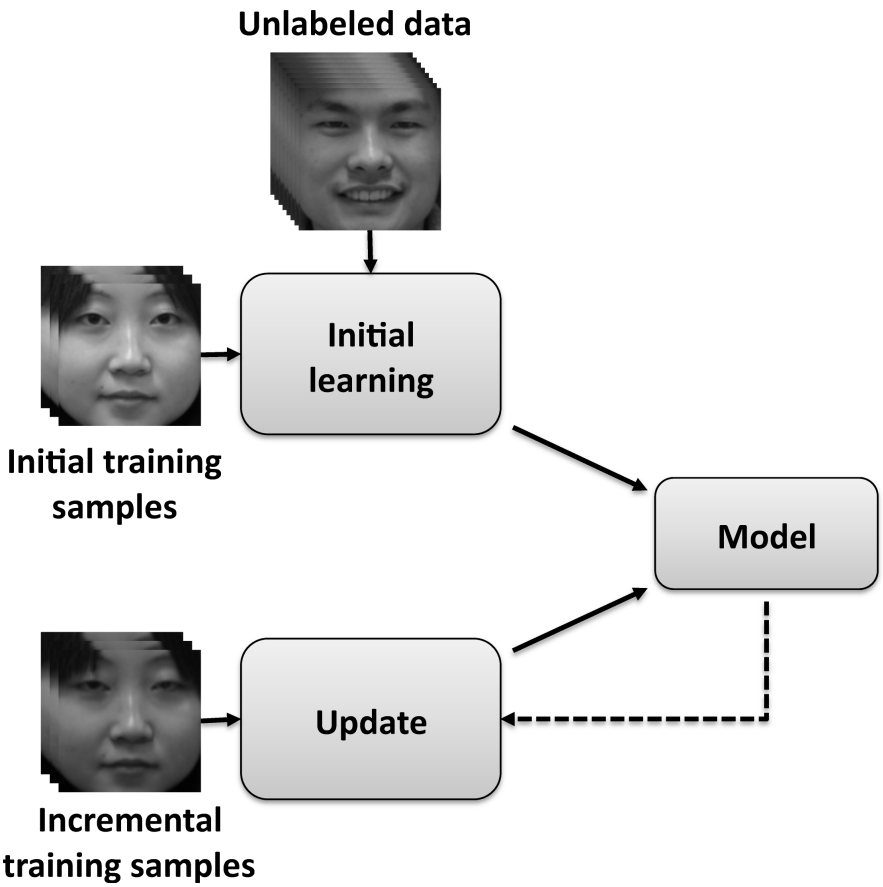


Figure 4-3: Block diagram of the evaluation protocol for face recognition experiments. At first, the model is learned using initial training samples and unlabeled data. With each incremental training batch the existing model is updated to obtain a new model.

data may arrive in batches of different sizes. For example, an incremental batch may consist of 2 new samples per class or it may contain 10 samples of two new classes. The incremental learning procedure utilizes the incremental batch and the existing learned model, and returns the model Υ_3 that incorporates all the new samples. Further, when another incremental batch arrives, the Υ_3 is used as the current existing model.

The proposed algorithm is evaluated in context to face recognition. As shown in Figure 4-3, in the proposed incremental semi-supervised discriminant analysis, initially the classifier model is learned which consists of the eigenmodels of manifold regularized total scatter obtained from unlabeled data (Eq. 4.23) and the eigenmodel of between-class scatter obtained from labeled data. With every new incremental batch, the classifier model is updated by obtaining new eigenmodel of between-class scatter (Eq. 4.17-4.22) and updating the discriminating components (Eq. 4.24).

Algorithm 1 Proposed Incremental Semi-Supervised Discriminant Analysis (ISSDA)

procedure INITIAL TRAINING

Input: Data matrix $\mathbf{X} = [\mathbf{X}_1 \dots \mathbf{X}_c]$ where $\mathbf{X}_i$ is the set of samples belonging to i^{th} class, unlabeled set $\mathbf{X}^{(u)}$, regularization parameters β_1 and β_2.

1. Compute total scatter $\mathbf{S}_T^u$ (Eq. 4.5), and graph Laplacian $\mathbf{L}^{(u)}$ from unlabeled set $\mathbf{X}^{(u)}$.
2. Obtain eigenvectors $\mathbf{P}_1$ and eigenvalues $\mathbf{\Lambda}_1$ of $\left(\mathbf{S}_T^{(u)} + \beta_1 \mathbf{X}^{(u)} \mathbf{L}^{(u)} \mathbf{X}^{(u)^T} + \beta_2 \mathbf{I}\right)$
3. Obtain mean of training samples $\boldsymbol{\mu}_1$
4. Count number of samples M_1
5. Count number of samples in each class and arrange them in n_1
6. Obtain eigenvectors $\mathbf{Q}_1$ and eigenvalues $\mathbf{\Delta}_1$ of between-class scatter matrix as explained in Eq. 4.15-4.17
7. Obtain discriminating components $\mathbf{V}_1$ as explained in Eq. 4.24

Return: $\mathbf{\Upsilon}_1 = \{\mathbf{V}_1, \mathbf{P}_1, \mathbf{\Lambda}_1, \boldsymbol{\mu}_1, M_1, \mathbf{Q}_1, \mathbf{\Delta}_1, n_1\}$: Learned initial model

end procedure

procedure INCREMENTAL LEARNING:

Input: Incremental batch data matrix $\mathbf{X} = [\mathbf{X}_1 \dots \mathbf{X}_c]$ where $\mathbf{X}_i$ is the set of samples belonging to i^{th} class, Current existing model Υ_1

1. Update number of sample M_3 (Eq. 4.18), number of samples per class n_3 (Eq. 4.19), sample mean $\boldsymbol{\mu}_3$ (Eq. 4.20)
2. Obtain eigenvectors $\mathbf{Q}_2$ and eigenvalues $\mathbf{\Delta}_2$ of between-class scatter matrix of incremental batch as explained in Eq. 4.15-4.17
3. Obtain the set of orthonormal basis $\mathbf{\Psi}_3$ by applying orthormalization function on $[\mathbf{Q}_1, \mathbf{Q}_2, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]$.
4. Eigendecopmose $\mathbf{\Psi}_3^T \mathbf{S}_{B,3} \mathbf{\Psi}_3$ to obtain rotation matrix $\mathbf{R}$ (Eq. 4.17). Set $\mathbf{Q}_3 = \mathbf{\Psi_3 R}$.
5. Set $\mathbf{P} = \mathbf{P}_3$ and $\mathbf{\Lambda} = \mathbf{\Lambda}_3$.
6. Obtain discriminating components $\mathbf{V}_3$ as explained in Eq. 4.24

Return: $\mathbf{\Upsilon}_3 = \{\mathbf{V}_3, \mathbf{P}_3, \mathbf{\Lambda}_3, \boldsymbol{\mu}_3, M_3, \mathbf{Q}_3, \mathbf{\Delta}_3, n_3\}$: Updated model

end procedure

4.1.3 Time Complexity

Let M, D, and C be the number of data samples, dimensionality, and number of classes respectively ($M \geq C$). The computation of total scatter and between-class scatter matrices require $O(MD^2)$ and $O(CD^2)$ operations respectively. The inversion of $\mathbf{S}_T$ and multiplication of inverted matrix with $\mathbf{S}_B$ both require $O(D^3)$ operations. Finally, the eigenvalue decomposition requires $O(D^3)$ operations. Thus, the overall time complexity of finding discriminant components as per the objective function in Eq. 4.4 is $O(MD^2 + D^3)$. Moreover, following the trick performed by Turk and Pentland [14] it can be further reduced to $O(MD^2 + min(M, D)^3)$. If the number of samples in existing batch and incremental batch are M_1 and M_2 respectively and $M_3 = M_1 + M_2$. The time complexity of (non-incremental) linear discriminant analysis is $O(M_3D^2 + min(M_3, D)^3)$. The computation of incremental linear discriminant analysis [207], [208] is $O(d_{T,1}^3 + d_{B,1}^3 + Dd_{T,3}d_{B,3})$, under the assumption that $M_2 \ll M_1$, where $d_{T,i}$ and $d_{B,i}$ are the number of eigenvectors in eigenmodels of $\mathbf{S}_{T,i}$ and $\mathbf{S}_{B,i}$ respectively. However, without any inequality constraint assumption between M_1 and M_2, its time complexity is $O(d_{T,3}^3 + d_{B,3}^3 + Dd_{T,3}d_{B,3})$, where $d_{T,3}$ and $d_{B,3}$ are the number of eigenvectors in the updated model, with $d_{T,3} \leq d_{T,1}+d_{T,2}+1$ and $d_{B,3} \leq d_{B,1}+d_{B,2}+1$. It should be noted that usually $M_3 \gg d_{T,3} \geq d_{B,3}$. Therefore, it provides significant improvement over classical (non-incremental) LDA.

The proposed algorithm aims at further reducing the computational complexity by updating only $\mathbf{S}_B$ incrementally. As the eigenmodel of total scatter is not to be updated, the time complexity of the proposed incremental learning is $O(d_{B,3}^3 + Dd_{T,3}d_{B,3})$. The additional learning from the unlabeled data is $O(kDM_u^2 + M_uD^2)$, where the first term corresponds to finding k-nearest neighbor (considering Euclidean distances) and the second term corresponds to matrix multiplication $\mathbf{XLX}^T$. M_u is the number of unlabeled samples. However, these additional operations for finding graph Laplacian are to be performed only once during offline learning from initial batch. The computational complexities of various approaches is summarized in Table 4.2.

4.2 Experiments and Results

Using face recognition application, the effectiveness of the proposed ISSDA algorithm is evaluated in terms of recognition performance (accuracy, training time, and consistency) along with its sen-

Table 4.2: Computational complexity analysis. M and D represent the number of samples and feature dimensionality. $d_{T,i}$ and $d_{B,i}$ is the number of components preserved in eigenmodels of total and between-class scatter matrices of i^{th} batch. M_u is the number of samples in unlabeled set and k is the neighborhood parameter of learning graph laplacian.

Algorithm	Time complexity
PCA	$O(min(MD^2 + D^3))$
LDA	$O(MD^2 + \min(M, D)^3)$
IPCA [202]	$O((d_{T,1} + d_{T,2} + 1)^3)$
ILDA [207], [208]	$O(d_{T,3}^3 + d_{B,3}^3 + Dd_{T,3}d_{B,3})$
SSDA [195]	$O(MD^2 + \min(M, D)^3 + kDM_u^2 + M_uD^2)$
ISSDA(Proposed)	$O(d_{B,3}^3 + Dd_{T,3}d_{B,3})$

sitivity to size of unlabeled set and regularization. Further, results are evaluated when new classes are incrementally added.

4.2.1 Database, Experiment Design, and Protocols

The experiments are performed on three face databases namely, (1) CMU-PIE [55], (2) NIR-VIS-2.0 [228], and (3) CMU-MultiPIE [61]. Figure 4-4 shows sample images from all three databases.

- The CMU-PIE face dataset contains 42,368 face images pertaining to 68 subjects with variations in pose, illuminations and expression. In our experiments, frontal face images with illumination changes (C27) are used which result in 42 face images per subject and a total of 2,856 images (i.e. 42×68).

- NIR-VIS-2.0 [228] face dataset consists of 5,093 visible and 12,487 near-infrared (NIR) spectrum frontal face images pertaining to 725 subjects captured in 4 sessions. In NIR spectrum, there are 5 to 50 images per subject, with a median of 18 images per person. In fact, there is only one subject with only 5 images. In VIS spectrum, there are 1 to 22 images per subject, with a median of 6 images per person. The experiments are performed for each spectrum individually in which cropped face images provided along with the dataset are used.[1]

[1]Since the proposed approach is not designed for heterogeneous face matching, a custom defined protocol is utilized instead of the the predefined protocol for the NIR-VIS-2.0 dataset.

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

Figure 4-4: Sample images from the (a) CMU-PIE dataset [55] (b) visible spectrum and (c) NIR images from VIS-NIR-2.0 dataset [228], and (d) CMU-MultiPIE dataset [61].

Table 4.3: Experimental protocols

Database	Number of Subjects			Number of Images	
	Total	Unlabeled	Labeled	Total	Per Subject
CMU-PIE [55]	68	30	38	2,856	42
NIR-VIS-2.0 [228] (VIS only)	725	300	425	5,093	1-22
NIR-VIS-2.0 [228] (NIR only)	725	300	425	12,487	5-10
CMU-MultiPIE [61]	337	150	187	32,780	20-140

- CMU-MultiPIE face dataset contains more than 7,50,000 images pertaining to 337 subjects, with 15 pose, 20 illumination and 6 expression variations captured across 4 sessions. For this research, we utilize 32,780 images pertaining to frontal pose (camera 05_1), neutral and smile expressions, and all the illumination variations.

In all the experiments, size of registered face image is set to 32 $\times$ 32 pixels, with 256 grey levels per pixel. All the following experiments are performed on raw pixel intensity features. We perform four sets of experiments,

1. to study the performance of the proposed approach with respect to batch (PCA, LDA, and SSDA) and other incremental learning (IPCA, ILDA) approaches.

2. to study effect of manifold regularization.

3. to study the effect of size of unlabeled set.

4. to study the effect of incremental addition of classes.

For each experiment, four sub-experiments are performed using the four datasets:1) CMU-PIE, 2) VIS images of NIR-VIS-2.0, 3) NIR images of NIR-VIS-2.0, and 4) CMU-MultiPIE datasets.

For all the datasets, the labeled and unlabeled sets are designed such that they are non-overlapping in terms of subjects. Details of both labeled and unlabeled splits are given in Table 4.3. In all the experiments, unlabeled set consists of images pertaining to subjects selected for unlabeled training. Note that for experiments pertaining to NIR images of NIR-VIS-2.0 dataset, the subjects with less than five images are made part of the unlabeled set. Labeled set is split into train and test sets.

Train set is further divided into training batches. For labeled data, the initial training batch consists of one labeled image per subject. Each incremental batch contains an additional labeled image per subject for training in CMU-PIE and NIR-VIS-2.0 dataset; whereas for CMU-MultiPIE dataset five additional images per subject are used in each incremental batch. In sub-experiments pertaining to CMU-PIE and CMU-MultiPIE, four such incremental training batches are formed whereas in sub-experiment pertaining to NIR-VIS-2.0 dataset, three such incremental training batches are formed. At all the stages, testing is performed on the predefined test set, which is non-overlapping with the train set. The results are reported with 15 times repeated random sub-sampling based cross validations.

4.2.2 Experiment 1: Comparative Evaluation

In this experiment, performance of the proposed ISSDA is evaluated and comparison is performed with

- three batch learning approaches:
  - Principal Component Analysis (PCA) [14],
  - Regularized Linear Discriminant Analysis (LDA) [158], and
  - Semi-supervised discriminant analysis (SSDA) [195], and
- two incremental approaches:
  - Incremental PCA (IPCA) [202] and
  - Incremental LDA (ILDA) [207], [208].

PCA is applied in two modes: one by augmenting the unlabeled training data to learn eigenspace (represented as PCA*[1]) and other by not augmenting (represented as PCA*). The PCA implementation of statistical toolbox of Matlab is utilized. Publicly available implementation of LDA[2], SSDA[2], and ILDA[3] are used. IPCA code is derived based on ILDA source code. The proposed semi-supervised incremental learning is evaluated in terms of accuracy, consistency, and time.

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/DimensionReduction.html
[3] http://www.iis.ee.ic.ac.uk/icvl/code.htm

Ideally, the incremental learning algorithm is expected to have a similar performance as the corresponding batch algorithm and should consume significantly less time. The *consistency measure* is to evaluate the similarity of the models obtained by incremental learning and batch learning.

Accuracy and Consistency

Tables 4.4 and 4.5 show the rank-1 identification accuracies of various algorithms with different incremental batches. Figure 4-6 shows the cumulative match characteristics (CMC) curves of the proposed algorithm and Table 4.6 shows the results pertaining to consistency evaluation. The consistency evaluation enables us to understand *how similar are the predicted labels of incremental and batch learning models*. The key observations are as follows:

- It can be observed that the proposed ISSDA, generally, yields similar identification performance compared to SSDA. As mentioned previously, incremental version of batch mode learning has minor effect on accuracy but significantly improves the computational time (discussed later). Further, it is also observed that the performance difference between LDA and ILDA is higher compared to SSDA and the proposed ISSDA. This performance improvement of ISSDA compared to ILDA (or IPCA) may be attributed to the graph Laplacian based regularization in the initial batch and incorporating learning with unlabeled data. As show in Figure 4-5, for most of the cases, the following trend is observed in terms of accuracy performance:

  IPCA$\approx$PCA* $\approx$ PCA $<$ ILDA $<$ {LDA, ISSDA} $<$ SSDA

- In the initial batch, the proposed ISSDA and SSDA do not yield exact same accuracy due to the difference in objective function of both the algorithms. Similar observations can be made for the case of PCA and LDA. The proposed approach computes the discriminant components by using eigenmodels of $\mathbf{S}_B$ and $\mathbf{S}_T$ (Eq. 4.24). This procedure is not exactly same as finding projection directions by solving generalized eigenvalue problem as in SSDA.

- The algorithms that utilize manifold assumption, i.e. SSDA and the proposed one, generally, have better accuracies than other DA algorithms the earlier batches. This suggests that SSDA and the proposed algorithm are less affected by the small sample size problem than other DA approaches.

Table 4.4: Rank-1 identification accuracy (mean±std-dev %) and computation time for sub-experiments pertaining to the CMU-PIE, and NIR-VIS-2.0 (VIS spectrum). Initial batch consists of one image per subject and each incremental batch consists of one new image per subject. Therefore, there are n images in initial and each incremental batch, where n is the number of subjects in the test split. *Algorithms are not incremental, therefore batch mode training results are reported.

Algorithm	**Accuracy**					**Time**				
	Initial batch	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 4**	**Initial batch**	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 4**
PCA*	36.8±2.1	57.0±3.5	68.3±2.9	75.8±2.7	81.2±2.2	**0.004±0.00**	**0.01±0.01**	**0.02±0.00**	**0.03±0.01**	**0.05±0.01**
PCA*[1]	31.0±1.7	50.1±2.7	62.7±2.8	71.2±2.8	77.9±2.5	2.56±0.24	2.59±0.24	2.60±0.25	2.71±0.15	2.65±0.12
LDA*	**38.1±2.0**	**93.9±2.0**	**98.1±0.6**	**99.4±0.3**	**99.7±0.1**	0.54±0.04	0.54±0.03	0.53±0.05	0.54±0.05	0.52±0.05
SSDA* [195]	75.1±4.2	95.7±1.4	98.3±0.5	99.3±0.3	99.6±0.1	1.23±0.13	1.21±0.06	1.24±0.04	1.28±0.08	1.33±0.06
IPCA [202]	36.8±2.1	57.4±3.4	69.0±2.9	76.6±2.6	82.0±2.1	**0.002±0.00**	**0.01±0.00**	**0.01±0.00**	**0.02±0.00**	**0.03±0.01**
ILDA [207], [208]	83.8±2.5	96.6±1.9	**98.7±0.8**	**99.7±0.2**	**99.9±0.1**	0.01±0.01	0.06±0.01	0.08±0.01	0.11±0.01	0.15±0.01
ISSDA	**91.0±1.8**	**97.1±1.2**	98.6±0.6	99.2±0.5	99.5±0.3	4.03±0.16	0.15±0.01	0.14±0.01	0.14±0.01	0.15±0.01

(a) CMU-PIE dataset.

Algorithm	**Accuracy**				**Time**			
	Initial batch	**Batch 1**	**Batch 2**	**Batch 3**	**Initial batch**	**Batch 1**	**Batch 2**	**Batch 3**
PCA*	87.7±1.1	94.3±1.1	96.8±0.8	97.9±0.5	**0.21±0.02**	**0.63±0.09**	**0.86±0.08**	**0.89±0.07**
PCA*[1]	87.7±1.1	94.3±1.1	96.8±0.8	97.9±0.5	0.96±0.11	1.03±0.10	1.07±0.11	1.07±0.17
LDA*	87.9±1.1	95.4±1.0	**97.6±0.8**	**98.6±0.5**	7.81±3.38	7.62±3.09	7.57±3.04	7.51±2.98
SSDA* [195]	**89.5±1.1**	**95.5±1.0**	**97.6±0.8**	**98.6±0.5**	6.52±0.16	7.05±0.19	7.55±0.20	7.72±0.22
IPCA	87.8±1.3	94.1±0.7	96.7±0.8	98.0±0.5	**0.16±0.01**	**0.37±0.03**	**0.58±0.05**	**0.78±0.02**
ILDA [207], [208]	85.8±1.3	91.9±1.0	95.4±0.6	97.3±0.3	0.41±0.04	2.46±0.14	3.27±0.23	3.77±0.15
ISSDA	**89.2±1.2**	**95.3±1.0**	**97.5±0.8**	**98.5±0.5**	3.92±0.17	0.98±0.05	1.12±0.09	1.18±0.09

(b) VIS images of NIR-VIS-2.0 dataset.

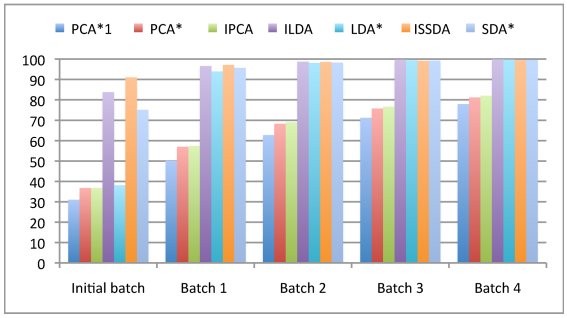
Table 4.5: Rank-1 identification accuracy (mean±std-dev %) and computation time for sub-experiments pertaining to the NIR-VIS-2.0 (NIR spectrum) and CMU-MultiPIE face datasets. Initial batch consists of one image per subject and each incremental batch consists of one new image per subject. Therefore, there are n images in initial and each incremental batch, where n is the number of subjects in the test split. *Algorithms are not incremental, therefore batch mode training results are reported.

Algorithm	**Accuracy**				**Time**			
	Initial batch	**Batch 1**	**Batch 2**	**Batch 3**	**Initial batch**	**Batch 1**	**Batch 2**	**Batch 3**
PCA*	51.6±1.7	65.7±1.2	73.1±0.6	77.5±0.6	**0.13±0.00**	**0.46±0.01**	**0.59±0.03**	**0.63±0.03**
PCA*[1]	51.9±1.7	65.9±1.2	73.2±0.7	77.5±0.6	0.86±0.05	0.93±0.05	0.94±0.04	0.96±0.06
LDA*	52.1±1.7	**81.4±0.9**	**88.8±0.5**	**91.7±0.5**	5.26±0.17	5.28±0.18	5.29±0.18	5.24±0.18
SSDA* [195]	**56.9±1.8**	78.9±0.9	86.9±0.6	90.5±0.6	10.06±0.25	10.64±0.22	11.37±0.33	11.97±0.23
IPCA	51.6±1.7	65.9±1.2	73.3±0.6	77.6±0.6	**0.17±0.01**	**0.36±0.04**	**0.55±0.01**	**0.72±0.02**
ILDA [207], [208]	**62.6±1.6**	70.9±0.9	76.3±0.6	79.7±0.6	0.43±0.01	2.54±0.17	3.36±0.17	3.87±0.22
ISSDA	61.7±1.7	**75.6±1.0**	**81.8±0.5**	**85.0±0.6**	3.77±0.11	0.98±0.05	1.14±0.05	1.17±0.09

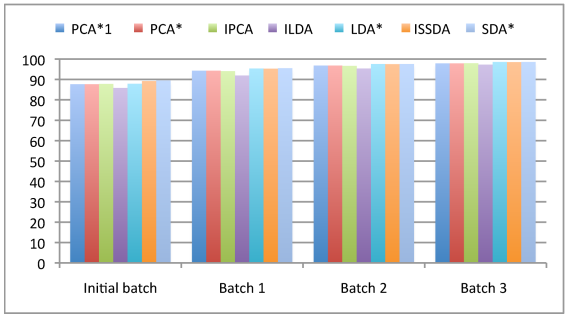
(a) NIR images of NIR-VIS-2.0 dataset.

Algorithm	**Accuracy**					**Time**				
	Initial batch	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 4**	**Initial batch**	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 4**
PCA*	50.4±1.0	68.4±1.2	77.9±1.1	83.8±0.8	87.8±0.6	**0.63±0.11**	**0.66±0.04**	**0.72±0.03**	**0.80±0.04**	**0.93±0.17**
PCA*[1]	50.3±1.0	68.2±1.3	77.7±1.1	83.7±0.8	87.7±0.6	1.71±0.37	1.71±0.21	1.80±0.31	1.77±0.12	1.89±0.14
LDA*	94.3±0.6	**99.0±0.2**	**99.7±0.1**	**99.8±0.0**	**99.9±0.0**	2.96±0.62	2.98±0.55	3.06±0.41	3.17±0.55	3.19±0.57
SSDA	**93.9±0.6**	98.9±0.2	99.6±0.1	**99.8±0.0**	**99.9±0.0**	30.20±1.00	34.39±3.40	37.87±1.42	42.38±1.70	47.03±1.79
IPCA	50.4±1.0	68.7±1.2	78.3±1.0	84.2±0.7	88.2±0.5	**1.05±0.10**	1.47±0.14	1.81±0.18	2.23±0.24	2.44±0.18
ILDA [207], [208]	**89.0±0.6**	76.9±1.1	82.4±0.8	88.1±0.6	91.7±0.6	1.13±0.38	1.99±0.48	2.46±0.56	2.85±0.37	3.15±0.62
ISSDA	88.4±0.9	**95.9±0.4**	**98.0±0.3**	**98.9±0.1**	**99.4±0.0**	33.85±3.12	**0.73±0.32**	**0.79±0.48**	**0.74±0.17**	**0.82±0.26**

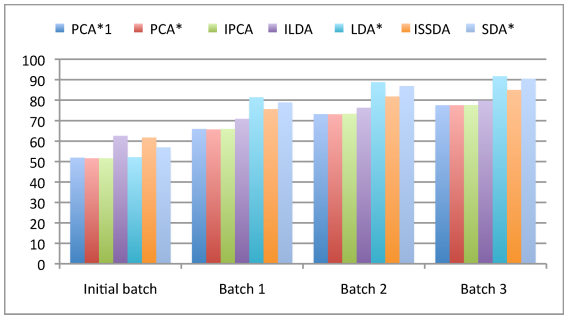
(b) CMU-MultiPIE dataset.



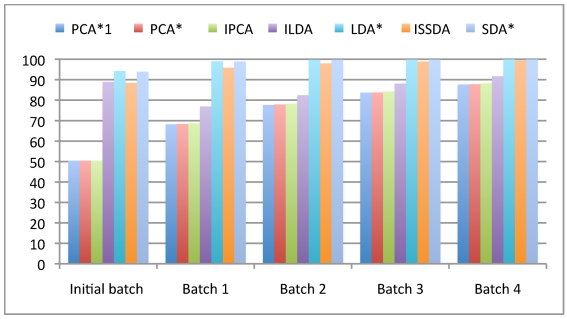
(a) CMU-PIE



(b) NIR-VIS-2.0 (VIS only)



(c) NIR-VIS-2.0 (NIR only)



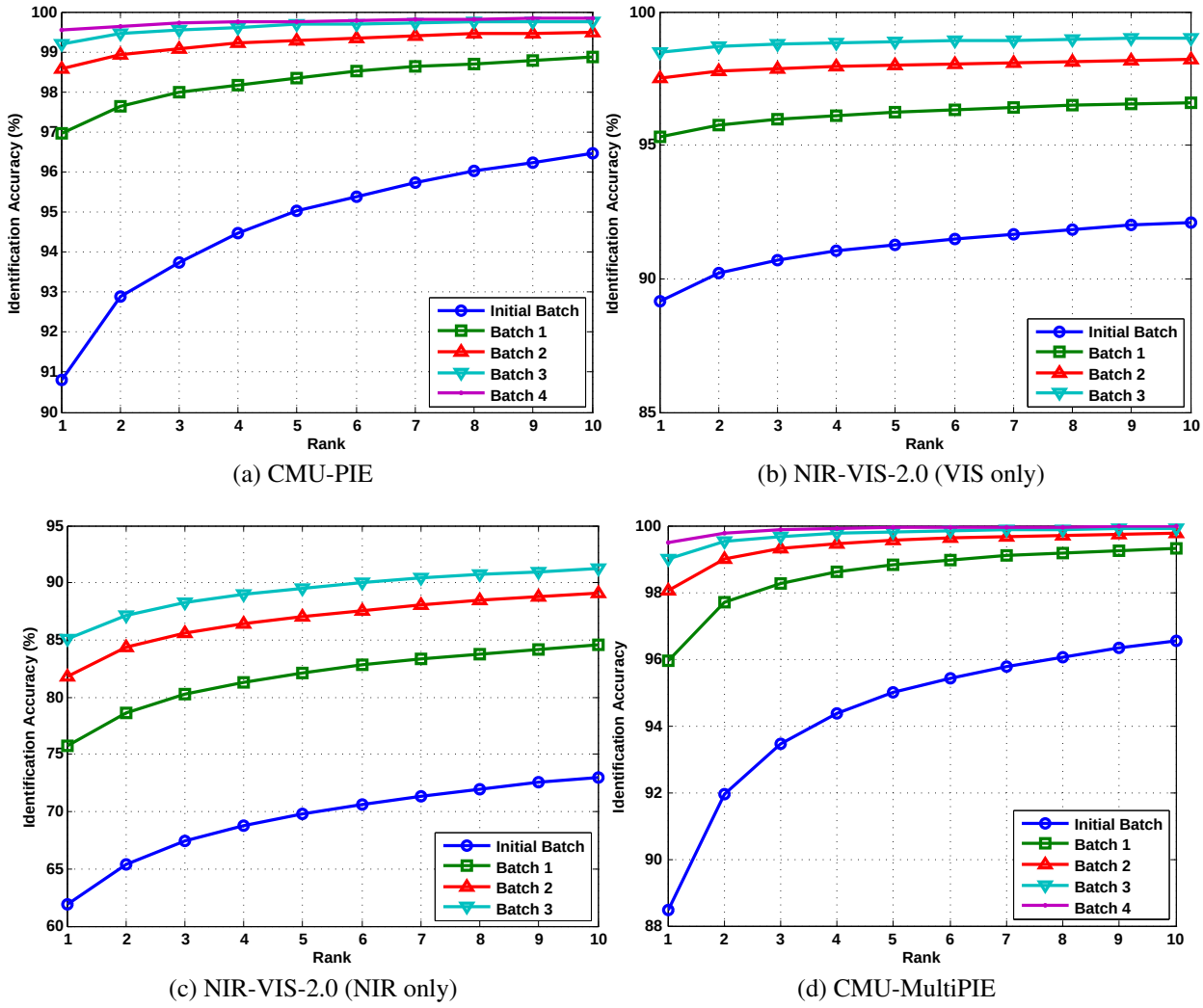
(d) CMU-MultiPIE

Figure 4-5: Rank-1 identification accuracy for sub-experiments pertaining to CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE face dataset. The graph representation of the accuracy helps understand the general trend among approaches.

- As shown in Table 4.6, to evaluate the consistency between SSDA and the proposed ISSDA, we utilize the confusion matrix representing similarity in class label predictions after arrival of the final batches. If two classification models are exactly same, the confusion matrix should have zero entries on minor diagonal. The results reported in Table 4.6 show that for three databases used (CMU-PIE, NIR-VIS-2.0 (VIS only), and CMU-MultiPIE), more than 99.5% and for NIR-VIS-2.0 (NIR only) database approximately 95% of the label predictions by the proposed ISSDA have agreed with SSDA, respectively.

Time

The time required for training different models is reported in Tables 4.4 and 4.5. The tables show the time required for computing the model on arrival of new batches. For the incremental approaches, i.e. ILDA and the proposed ISSDA algorithm, the time required for updating the existing



(a) CMU-PIE

(b) NIR-VIS-2.0 (VIS only)

(c) NIR-VIS-2.0 (NIR only)

(d) CMU-MultiPIE

Figure 4-6: CMC curves of the proposed approach for (a) CMU-PIE, (b) NIR-VIS-2.0 (VIS only), (c) NIR-VIS-2.0 (NIR only), and (d) CMU-MultiPIE datasets. Consistently, the incremental update improves the identification performance.

model are shown. On the other hand, for the batch approaches, i.e. PCA, LDA and SSDA, the time required for learning a new model using all the available training samples available till the current batch is shown. The models are computed with Matlab R2010 on a machine with two Intel Xeon E5640 (2.67GHz) processors and 48GB RAM. The key observations are as follows:

- For the proposed algorithm, the time required in learning from the initial batch is higher, however it is very small for all the incremental batches. Note that in the purview of incremental learning the one time high computation cost of initial training is acceptable.
- With the experiments on bigger data (Table 4.4(b) and 4.5(a)), the time difference between SSDA and the proposed ISSDA becomes more apparent. It can be seen that updating the existing model requires less than approximately $1/10^{\text{th}}$ and one $1/50^{\text{th}}$ of the time than learning a new model in the batch-mode for NIR-VIS-2.0 and CMU-MultiPIE datasets, respectively.
- Also note that, to incorporate new samples in non-incremental approaches, the new model is learned and the old model has to be discarded. From systems perspective, this changes in the classification model may result in significantly more downtime of system compared to the proposed incremental approach.
- In the proposed ISSDA, S_T is estimated using the large unsupervised training data and then not updated using incremental training samples. In order to understand the effect of updating S_T with incremental training data, we have compared the performance of ISSDA with and without S_T update. It is observed that utilizing the unlabeled data to incrementally update total scatter estimate yields $0 - 0.2\%$ accuracy improvement while increasing the computation times by $2 - 3$ times compared to the proposed ISSDA.

4.2.3 Experiment 2: Effect of Manifold Regularization

The objective function in the proposed approach (Eq. 4.12) has a term representing manifold regularization. To evaluate the impact of the manifold regularization, all the four experiments are also performed without considering it in the objective function. In other words, the parameter β_1 is set to zero to remove manifold regularization from the proposed approach. The results are shown in Table 4.7. The key observations from this set of experiments are as follows:

Table 4.6: Confusion matrix for comparing the performance of SSDA and ISSDA. ✓ and ✗ represent the percentage of correctly classified and misclassified samples respectively.

Confusion matrix @ Rank 1		SSDA	✓	✗
CMU-PIE	Proposed	✓	99.71	0.00
		✗	0.07	0.22
NIR-VIS-2.0 (VIS only)	Proposed	✓	98.14	0.07
		✗	0.07	1.72
NIR-VIS-2.0 (NIR only)	Proposed	✓	84.45	0.14
		✗	5.88	9.53
CMU-MultiPIE	Proposed	✓	99.42	0.00
		✗	0.53	0.04

- Comparison with Table 4.4 and 4.5 shows that in all the experiment pertaining to CMU-PIE dataset, the performance is superior when manifold regularization is used, whereas, in the other three datasets performance difference is not statistically significant.

- The limited amount of training data with respect to the amount of variations in the datasets, leads to performance deterioration in absence of manifold regularization. CMU-PIE, which has the most variations because of large illumination changes, is affected the most in absence of regularization. The effect of manifold regularization is not significant in other three experiments. The deteriorating effect on CMU-PIE may be attributed to over-fitting, as the learned subspace would try to fit only the variations which are seen in the limited training set. Incorporating manifold regularization helps to address the issue of over-fitting.

4.2.4 Experiment 3: Effect of Size of Unlabeled Set

Eq. 4.12 shows that the estimation of total scatter is dependent on the size of the unlabeled set. To understand how the size of unlabeled set affects the overall identification accuracy, performance is measured by varying the size of unlabeled set. Size of unlabeled set is varied in terms of number of subjects used to constitute the set. Results pertaining to this experiment are reported in Table 4.8. The key observations of this set of experiments are as follows:

Table 4.7: Rank-1 identification accuracy (mean±std-dev %) on the CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE dataset without using manifold regularization. The experiment design and protocol is same as Experiment 1 except that $\beta_1 = 0$ is set. For easier comparison, the results obtained with manifold regularization are shown reported within brackets. The smaller the dataset the more noticeable is the performance drop when manifold regularization is not used.

Datasets	**Regularizer**	**CMU-PIE**	**VIS only**	**NIR only**	**CMU-MultiPIE**
Initial batch	without	40.5±2.9	86.9±1.1	61.4±1.8	88.6±0.9
	with	91.0±1.8	89.2±1.2	61.7±1.7	88.4±0.9
Batch 1	without	59.1±3.7	93.8±1.2	75.4±1.0	96.0±0.4
	with	97.1±1.2	95.3±1.0	75.6±1.0	95.9±0.4
Batch 2	without	69.7±3.1	96.5±0.9	81.6±0.5	98.0±0.3
	with	98.6±0.6	97.5±0.8	81.8±0.5	98.0±0.3
Batch 3	without	76.2±2.8	97.7±0.7	84.9±0.5	99.0±0.1
	with	99.2±0.5	98.5±0.5	85.0±0.6	98.9±0.1
Batch 4	without	81.1±2.0	-	-	99.4±0.0
	with	99.5±0.3	-	-	99.4±0.0

- On the CMU-PIE dataset a significant performance improvement is obtained by increasing the size of the unlabeled set. However that is not the case with the rest of the experiments, where the performance difference of 1-4% is observed with variation in unlabeled set size. Since the training set of CMU-PIE contains images pertaining to small number of subjects (only 38), the total scatter estimate may be inaccurate. Moreover, the dataset contains images with well structured variations. Therefore, adding new subjects in unlabeled set should be significantly affecting the total scatter estimate. While this may not be the case with NIR-VIS-2.0 and CMU-MultiPIE datasets, since they not contain face images with well structured variations and face images pertaining to limited number of subject, respectively. Thus, it is observed that *addition of more images is helpful, only if large (inter-class and intra-class) variations are captured by these images*.

- We further observe that the improvement in results of initial batches is more compared to successive incremental batches. This shows that role of unlabeled set is more important with small training set. In a way, the small size of training dataset is being compensated by a larger unlabeled set.

Table 4.8: Rank-1 identification accuracy (mean±std-dev %) on the CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE datasets with varying sizes of unlabeled set. It can be observed that moving from left to right in a row (increasing the size of unlabeled set), and top to bottom (updating the model with incremental batches) yields better accuracy.

No. of subjects in unlabeled set	5	10	15	20	25	30
Initial batch	61.2±6.0	70.5±3.7	75.6±3.2	78.6±3.2	90.9±1.8	91.0±1.8
Batch 1	79.8±5.2	87.6±3.1	90.5±2.5	92.2±2.0	96.9±1.3	97.1±1.2
Batch 2	87.0±3.5	92.8±1.2	94.8±1.1	96.0±1.0	98.5±0.6	98.6±0.6
Batch 3	91.2±2.5	95.8±0.9	97.0±1.0	97.6±0.7	99.2±0.5	99.2±0.5
Batch 4	93.6±2.0	97.2±0.6	98.0±0.6	98.4±0.4	99.5±0.3	99.5±0.3

(a) CMU-PIE

No. of subjects in unlabeled set	10	20	30	40	50	100	150	200	250	300
Initial batch	85.5±2.2	88.1±1.2	88.5±0.9	88.9±1.0	88.9±1.0	89.5±1.0	87.9±2.4	89.5±0.9	89.4±1.3	89.2±1.2
Batch 1	92.9±1.9	94.3±1.1	94.7±1.1	94.9±1.1	94.8±1.1	95.1±1.1	94.4±1.6	95.2±0.8	95.3±1.0	95.3±1.0
Batch 2	95.8±1.3	96.8±0.8	97.1±0.8	97.2±0.7	97.1±0.7	97.3±0.7	96.9±0.9	97.5±0.7	97.5±0.8	97.5±0.8
Batch 3	97.2±0.9	97.9±0.6	98.1±0.6	98.2±0.5	98.2±0.5	98.3±0.6	98.2±0.5	98.5±0.5	98.5±0.5	98.5±0.5

(b) NIR-VIS-2.0 (VIS only)

No. of subjects in unlabeled set	10	20	30	40	50	100	150	200	250	300
Initial batch	58.8±2.4	60.7±2.0	61.1±1.8	61.0±1.9	61.2±1.8	61.1±1.8	60.9±1.9	61.3±1.8	61.6±1.7	61.7±1.7
Batch 1	72.7±1.8	74.6±1.3	75.0±1.1	74.9±1.1	75.0±1.1	74.9±1.0	74.9±1.0	75.3±1.0	75.5±1.1	75.6±1.0
Batch 2	79.1±1.2	81.0±0.8	81.3±0.6	81.3±0.5	81.4±0.7	81.3±0.6	81.2±0.5	81.5±0.4	81.7±0.5	81.8±0.5
Batch 3	82.6±1.2	84.3±0.7	84.6±0.6	84.6±0.5	84.6±0.7	84.5±0.6	84.4±0.6	84.7±0.5	84.9±0.6	85.0±0.6

(c) NIR-VIS-2.0 (NIR only)

No. of subjects in unlabeled set	10	20	30	40	50	75	100	125	150
Initial Batch	89.0±0.8	89.3±1.0	89.3±0.8	89.0±1.0	89.0±0.8	88.7±1.0	88.7±1.0	88.4±1.0	88.4±0.9
Batch 1	96.2±0.3	96.3±0.6	96.2±0.4	96.2±0.4	96.1±0.4	96.1±0.4	96.0±0.4	95.9±0.5	95.9±0.4
Batch 2	98.2±0.2	98.2±0.3	98.2±0.2	98.2±0.2	98.1±0.2	98.1±0.3	98.1±0.3	98.0±0.2	98.0±0.3
Batch 3	99.0±0.1	99.1±0.1	99.0±0.1	99.0±0.1	99.0±0.1	99.0±0.1	99.0±0.1	99.0±0.1	98.9±0.1
Batch 4	99.4±0.1	99.5±0.1	99.5±0.1	99.5±0.1	99.5±0.0	99.4±0.0	99.5±0.0	99.4±0.0	99.4±0.0

(d) CMU-MultiPIE

Table 4.9: Rank-1 identification accuracy (mean±std-dev %) on the CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE datasets by incrementally adding new subjects. The accuracy difference when incrementally adding new subjects (classes) is similar to that of incremental addition of new images of existing subjects as reported in Table 4.4 and 4.5.

No. of subjects in incremental batch	10	10	10	8
Proposed ISSDA	99.6±0.4	99.0±0.6	98.5±0.5	97.4±1.0
SSDA	99.5±0.6	99.5±0.4	99.5±0.3	99.5±0.3

(a) CMU-PIE

No. of subjects in incremental batch	105	105	105	110
Proposed ISSDA	98.1±1.3	98.3±0.6	98.2±0.5	98.3±0.5
SSDA	98.3±1.2	98.4±0.5	98.4±0.4	98.5±0.4

(b) NIR-VIS-2.0 (VIS only)

No. of subjects in incremental batch	105	105	105	110
Proposed ISSDA	87.9±1.9	86.2±1.0	85.7±1.0	85.3±0.9
SSDA	91.0±1.5	90.4±0.8	90.2±0.9	90.4±0.7

(c) NIR-VIS-2.0 (NIR only)

No. of subjects in incremental batch	30	30	30	30	30	37
Proposed ISSDA	99.9±0.1	99.8±0.1	99.7±0.1	99.6±0.1	99.5±0.1	99.6±0.1
SSDA	100.0±0.0	99.9±0.0	99.9±0.0	99.9±0.0	99.9±0.0	99.9±0.0

(d) CMU-MultiPIE

4.2.5 Experiment 4: Incremental Addition of Classes

The results of Experiment 1 showcase that the proposed approach is able to maintain accuracy and consistency with significantly less time. However, the evaluation protocol of Experiment 1, 2 and 3 is designed such that incremental batches consist of images of *existing subjects* (i.e. sample/instance based incremental learning). Experiment 4 is performed to evaluate the performance when images of *new subjects* are added incrementally (i.e. class based incremental learning). In this experiment, labeled training set is divided into four (for CMU-PIE and NIR-VIS-2.0 dataset) or five batches (for CMU-MuliPIE); where each batch contains images of new subjects. For each dataset, each batch consists of almost the same number of subjects. Labelled set of CMU-PIE and NIR-VIS-2.0 dataset consists of 5 images per subjects, whereas the same for CMU-MultiPIE is 25. The split of databases and the results are reported in Table 4.9. Note that, after incorporating new subjects in the classification model, the test set is also appended with the samples corresponding to novel subjects. Therefore, with every incremental training, the test set also increases. Owing to changing test set sizes, the performance of different batches should not be compared; instead the performance of different algorithms in the same batch should be compared. The key observations are as follows:

- In the experiment (Tables 4.9) the performance is affected very little ($0 - 2\%$) with incremental addition of subjects, which suggests that the classifier model is effectively updated with the samples of novel subjects.

- For CMU-PIE, NIR-VIS-2.0 (VIS only), and CMU-MultiPIE the accuracy of the proposed incremental semi-supervised discriminant analysis is comparable with SSDA. The addition of new subjects does not seem to have significant impact on accuracy.

- For the sub-experiment pertaining to NIR-VIS-2.0 (NIR only) the accuracy difference of 5-7% can be observed with respect to SSDA. The observation made in class based incremental learning is consistent with sample/instance based incremental learning (i.e. Experiment 1, Tables 4.4 and 4.5).

4.3 Summary

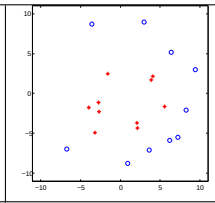
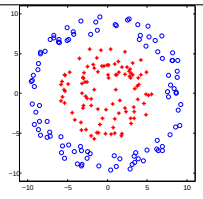
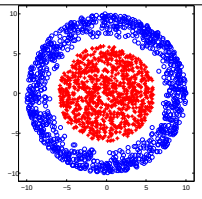
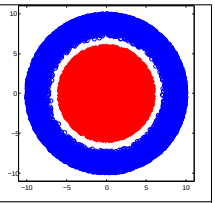
In this research, we propose an incremental semi-supervised discriminant analysis algorithm to mitigate the challenges such as batch learning and inability to utilize large unlabeled data which typically affect traditional discriminant analysis approaches. In the proposed algorithm, while between-class scatter is updated incrementally, total variability is estimated from large unlabeled training data and therefore does not require to update the total scatter with new increments. A case study in face recognition is presented with evaluations on the CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE datasets. The results show that the proposed incremental approach (1) has better identification accuracy than LDA and ILDA, (2) is consistent with the batch counterpart with lesser computational requirements, and (3) can effectively incorporate novel classes in incremental batches. Moreover, the study includes experiments examining the importance of manifold regularization and size of unlabeled set.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Large-scale Face Recognition by Leveraging Subclasses in Kernel SVM

Table 5.1: Training time as a function of the number of training instances for a synthetic two-dimensional dataset `two concentric circles` (`2CC`).

Visualization				
Number of instances	20	200	2000	20,000
Training time (seconds)	1.6×10^{-2}	3.6×10^{-1}	1.9×10^{1}	3.3×10^{3}

More and more applications are collecting large volumes of diverse data to be able to predict or make decisions. Currently, the size of the largest biometric data is more than a billion, the number of individuals with bank accounts has increased by more than 700 million in the last three years, and YouTube generates billions of views everyday. This has lead to an ever-growing popularity and importance of machine learning and pattern classification algorithms, particularly scalable machine learning models. Traditionally, the most important parameter in selecting a classification model is the accuracy of the classifier for a given problem; however, scalability of the classifier is now becoming another important factor.

A large number of classification techniques exist in the machine learning literature; each with their own advantages and limitations, based on their underlying assumptions. Support vector machine (SVM) [112] has been one of the widely used classification algorithms in a variety of domains and has shown excellent results in various applications including computer vision related problems (e.g. object classification [229] and pedestrian detection [169]). The efficiency of modeling decision boundary and, in turn, yielding impressive classification results have made SVMs desirable for real-world scenarios. Due to their impressive classification accuracies on various problems, one can potentially learn effective (in terms of classification accuracy) SVM models using large data. Numerous variants of SVM have been proposed in the literature to make large-scale training efficient [230]–[234]. However, there are two major limitations of SVM in context to large data.

- **Computational complexity:** The core optimization function of SVM is a quadratic programming (QP) problem. Therefore, the training time complexity of standard SVM is $O(n^3)$ [235], where n is the number of training instances.

- **Space complexity:** Training an SVM has space complexity of $O(n^2)$ [235]. This estimate is, typically, dominated by the space required for storing the kernel matrix.

To better understand the role of high time and space complexities, we show an example with a synthetic dataset termed as `two concentric circles` (`2CC`). As illustrated in Table 5.1, it is a 2D two-class dataset where the samples of each concentric circular band corresponds to one class. Computing the nonlinear hyperplane to separate the two concentric circles requires learning kernel SVM models. We perform experiments with varying number of data points in each circle and the results are summarized in Table 5.1[1]. The results show that depending on the number of points in the two circular bands, the training time of SVM changes significantly. With 20 data points, LibSVM requires 1.6×10^{-2} seconds whereas with 20,000 data points, i.e., increasing the number by three orders of magnitude, the training time increases to five orders of magnitude. This shows that traditional SVM solvers (here, LibSVM) are not optimized for large-scale learning.

The approaches proposed for scalable SVM can be grouped into four categories: (i) reduced training set size, (ii) incremental learning, (iii) improved solver, and (iv) leveraging hardware.

[1]The training time is computed with LibSVM.

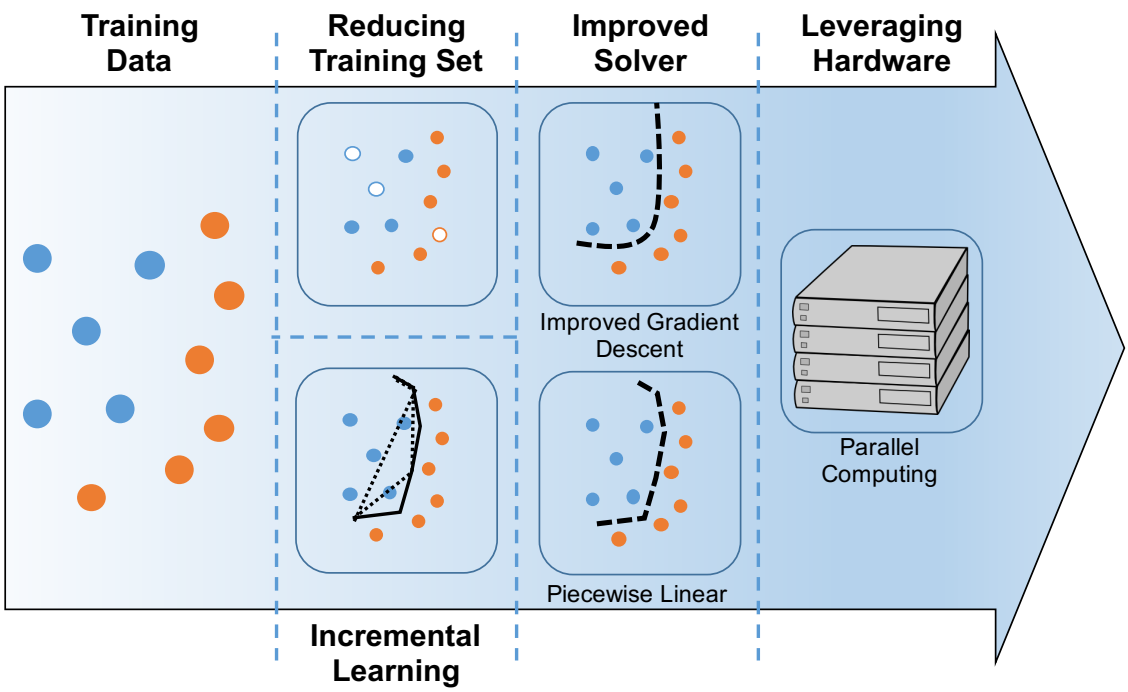


Figure 5-1: Illustrating four categories of approaches designed for scalable SVM learning.

As shown in Figure 5-1, these algorithms either operate at one of the steps involved in the SVM pipeline or incrementally update the learned model. We next review some of the algorithms in each of the four categories.

1. **Reducing Training Set Size:** The approaches that operate at data input stage, generally, propose to reduce the size of the training set by either dividing or reducing the training set into subsets. Since all the subsets are operated independently, the process is inherently parallelizable. Another advantage is that since these approaches operate at the first stage of training, the benefits of efficient reduction of the training set are observed at the solver and execution stages as well. From all the subsets, the required information is extracted, for instance, Lagrange multipliers and candidate support vectors. Later, the information from individual subsets is combined to obtain the final model.

   Among one of the first such approaches, Yu, Yang, and Han [236] proposed a top-down hierarchical clustering approach. The initial model is learned from the centroids of the top (biggest) clusters. In the subsequent stages, the model is updated based on the children

(smaller) clusters. A similar top-down approach is proposed by Boley and Cao [237]. Graf, Cosatto, Bottou, *et al.* [238] proposed Cascade SVM to learn models on disjoint subsets of the training set in parallel. The final SVM model is learned on a cumulative set of SVs obtained after iteratively processing the subsets.

Another set of techniques has focused on reducing the training set size in order to formulate scalable SVM models. In the ideal case, the reduced training set should consist of only those samples which are support vectors of the global solution. The reduced SVM and its variants [239], [240] include a candidate SV set selection stage followed by learning standard SVM model on it. Similarly, Ilayaraja, Neeba, and Jawahar [241] aimed at estimating concise candidate SV set in the multi-class scenario by exploiting the redundant nature of SVs amongst individual binary classifiers. Wang, Neskovic, and Cooper [242] explored the geometric interpretation of SVM to obtain the candidate SV set. Recently, Hsieh, Si, and Dhillon [243] proposed a divide-and-conquer SVM (DCSVM). Kernel k-means is first employed to divide the training set into subsets and a set of support vectors (SV) is obtained from each subset. The SVs are pooled and considered as the refined training set. Iteratively, the subsets are created using kernel k-means and the models are learned. The number of subsets is reduced in each subsequent iteration. DCSVM is currently one of the fastest SVM variants. Although not with the focus on scalability, Tong and Koller [244] proposed an active learning based approach to mitigate the need for large dataset.

2. **Incremental Learning:** A set of approaches inspired from incremental learning paradigm are also explored in literature. These approaches learn from incremental data streams and do not require to operate on the whole training set. This inherently results in reduced space requirements.

   Incremental SVM [245] variants have been introduced for more than two decades now. Since incremental SVM approaches do not require to keep the whole training set in the memory, their space complexity, typically is scalable for large training sets. Syed, Liu, and Sung [245] empirically showed that to incrementally update an existing SVM model, it is sufficient to learn a model from the combined pool of existing SVs and the SVs of the incremental batch. As an offshoot, it provides an empirical basis for utilizing SVs as the representative of the

decision boundary. “Incremental support vector machine learning: A local approach” [246] proposed using the locality information to update an SVM model with a new sample. Poggio and Cauwenberghs [247] provided a theoretical framework to increment or decrement the existing SVM model with a sample. Karasuyama and Takeuchi [248] extended the framework for incrementing existing SVM model with multiple samples.

3. **Improved Solver:** This category focuses on making the quadratic programming solver of SVM more efficient to handle large datasets. They can be grouped into either improving the gradient descent or obtaining the piece-wise linear solutions.

    - **Improved Gradient Descent:** One of the earliest research for addressing the computationally highly complex constrained QP focuses on reformulating the objective function in an unconstrained optimization function [249]. The proposed least square SVM classifier operates on the primal formulation by reformulating the optimization function into a set of linear equations. Other research efforts in similar directions are by Shalev-Shwartz, Singer, Srebro, *et al.* [233], Bottou and Lin [250], and Langford, Li, and Strehl [251] that use iterative algorithms such as stochastic gradient descent. Although extremely efficient for learning linear SVMs, the major limitation is that the approaches in this category may be difficult to apply with kernel SVMs due to their primal formulations and/or large kernel matrix computations.

    - **Piece-wise Linear Solutions:** These techniques operate by approximating the actual optimization problem. Such approaches focus on utilizing the intuition that even a nonlinear decision boundary is linear in small sections/local regions [252]. Huang, Mehrkanoon, and Suykens [253] proposed a piece-wise linear SVM approach via piece-wise linear feature mapping. Similarly, Fornoni, Caputo, and Orabona [254] proposed an approach that can leverage the piece-wise linear structure in the multiclass scenario with class specific weights. Ladicky and Torr [255] proposed to obtain local coding of each data point based on its local neighborhood. However, this approach is not aimed for large-scale learning. Kecman and Brooks [256] proposed to use the training samples in the vicinity of a query sample to obtain the final classification. Recently, Johnson and Guestrin [257] modeled the piecewise linearity property in terms of work-

ing set selection for improved scalability. It is to be noted that the locally linear SVM variants are not necessarily developed with the focus on large-scale learning. However, they provide the basis for utilizing the locally linear structure of complex decision boundaries for nonlinear classification.

4. **Leveraging Hardware:** This category is motivated by the availability of parallel computing hardware. The focus is to modify the solver algorithms for execution on multicore or multiprocessor environment. The research direction exploring the use of parallel processing and the hardware technology such as multicore processors [235] and distributed computing environments [258], [259] has resulted in various SVM variants for large-scale learning. Zanni, Serafini, and Zanghirati [260] proposed parallelization of stochastic gradient descent to exploit the multicore architecture of processors. Tsang, Kwok, and Cheung [235] proposed core vector machine that is specifically designed for utilizing multiple cores of processors. In order to efficiently leverage distributed and parallel processing environment, Do and Poulet [261] proposed a variant of least square SVM. Moreover, the inherently incremental nature of the approach makes its space complexity more suitable for large-scale learning. Caragea, Caragea, and Honavar [262] and Forero, Cano, and Giannakis [263] proposed approaches that rely on exchanging support vectors among sites (processing units) to learn the model in distributed computing environments. Do and Poulet [264] proposed to partition the training data and to learn parallel local SVM models on each of them. In a similar partitioning-based approach, Guo, Alham, Liu, *et al.* [265] proposed to leverage map-reduce framework for training SVM in heterogeneous parallel computation infrastructure.

Other approaches proposed for efficient large-scale learning include utilization of semi-supervised training data [266], leveraging the sparse nature of training data [267], and approximating the kernel equivalent high dimensional representation [268].

In this research, we propose between-subclass piece-wise linear solutions for large scale kernel SVM. The proposed Subclass Reduced Set (SRS) SVM takes advantage of the subset based approaches and piece-wise linear approaches. It focuses on splitting the nonlinear optimization problem into multiple linear optimization problems each operating on a significantly smaller fraction of the training data. Since the SVM solvers typically have super-linear time complexity, ap-

plying solver on such candidate set yields significant time improvements. This research proposes SRS-SVM which leverages subclass structures of data in order to reduce the time complexity of obtaining the decision boundary. We also provide a tree-like model of the proposed SRS-SVM, thereby extending it to its generalized hierarchical version, termed as Hierarchical SRS (HSRS)-SVM. Experiments are performed on four synthetic nonlinear datasets and six real-world datasets, namely `adult` [269], `IJCNN1` [270], `CIFAR-10` [271], forest cover (`covertype`) [272], face detection dataset from the Pascal Large Scale Learning Challenge (`LSL-FD`) [273], and Labeled Faces in the Wild (`LFW`) dataset [48]. The results are shown in comparison to LibSVM and state-of-the-art SVM variants proposed for large-scale data. The results demonstrate the effectiveness of the proposed approach on various datasets.

5.1 Preliminaries of SVM

This section briefly summarizes the basic formulation of support vector machine and defines some terms to facilitate explanation of the proposed approach.

SVM [112] is one of the widely used classification technique which falls under the category of discriminative classifiers. Let $\mathbf{x}_i$, $i = \{1, 2, \ldots, n\}$ be n training samples and $y_i = \pm 1$ be their corresponding class labels. A part of the objective of linear SVM is to obtain a projection direction $\mathbf{w}$ and a bias b such that samples of each class are on the different side of the separating plane, i.e.

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq +1, \forall y_i = +1, \text{and} \tag{5.1}$$

$$\mathbf{w} \cdot \mathbf{x_i} + b \leq -1, \forall y_i = -1 \tag{5.2}$$

Equivalently, this constraint can be written as $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, i \in \{1, 2, \ldots, n\}$. The parallel hyperplanes $\mathbf{w} \cdot \mathbf{x} + b = +1$ and $\mathbf{w} \cdot \mathbf{x} + b = -1$ are separated by a margin of width $\frac{2}{||\mathbf{w}||}$. The optimal $\mathbf{w}$ and b maximize the class separation by maximizing the margin. Thus, the optimization function of linear SVM becomes $\arg\max_{\mathbf{w},b} \frac{2}{||\mathbf{w}||}$, s.t. $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1$. Without changing the actual solution, the equivalent optimization function is

$$\arg\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2, \text{ s.t. } y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 \tag{5.3}$$

Practically useful formulation of SVM utilizes *soft margin* that tries to obtain as much cleaner decision boundary as possible. In other words, the model is allowed to misclassify training samples to a certain degree (represented by slack variable ξ_i), i.e. $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 - \xi_i$. Correspondingly, the optimization problem takes the form of

$$\arg\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i, \;\; \text{s.t.} \;\; y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 - \xi_i, \;\; \xi_i \geq 0 \tag{5.4}$$

where, C is the misclassification cost. Eq. 5.4 is called the *primal* form of the (soft-margin) SVM optimization function. By utilizing the Lagrangian multipliers α, the equivalent *dual form* of the optimization becomes

$$\arg\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}, \;\; \text{s.t.} \;\; 0 < \alpha_i \leq C \tag{5.5}$$

For complex, linearly non-separable datasets, a nonlinear decision boundary can be obtained by projecting the training samples in a higher dimensional space using transformation function $\phi(\cdot)$. To achieve this, $\mathbf{x}_i$ is replaced by $\phi(x_i)$ in Eq. 5.5. Further, by defining a kernel function $k(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i}) \cdot \phi(\mathbf{x_j})$, the optimization function of nonlinear SVM takes the form,

$$\arg\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x_i}, \mathbf{x_j}), \;\; \text{s.t.} \;\; 0 < \alpha_i \leq C \tag{5.6}$$

Having obtained the optimal multipliers α, the projection direction $\mathbf{w}$ is obtained as,

$$\mathbf{w} = \sum_{i} \alpha_i y_i \phi(\mathbf{x_i}) \tag{5.7}$$

As $\mathbf{w}$ and b define the hyperplane separating samples from two classes, and that $\mathbf{w}$ is defined as a linear summation of $\mathbf{x_i}$ makes it intuitive that only the α_i corresponding to the samples near the decision boundary are non-zero. Only these samples with non-zero multipliers, that contribute in defining $\mathbf{w}$, are called Support Vectors (SVs). All the points that are outside the margin get zero coefficient value assigned. In other words, $\alpha_i = 0, \; i \in \{j | y_j(\mathbf{w} \cdot \mathbf{x_j} + b) > 1\}$.

5.2 Reduced Set and Variants

We present the definitions and theorems associated to reduced set with respect to SVMs.

Definition 1 ***Reduced Set*** *(RS) is a subset of the training set indices. For a training set with n samples, the index set $T_{RS} \subset \{1, 2, \ldots, n\}$ defines a Reduced Set.*

Definition 2 ***Representative Reduced Set*** *(RRS) is a reduced set that yields the same decision boundary as the whole training set. Let α and $\hat{\alpha}$ represent the Lagrangian coefficients for the optimization functions of the whole training set and its reduced set T_{RS}, respectively. T_{RS} is a representative reduced set if $\sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i} = \sum_{j \in T_{RS}} \hat{\alpha}_j y_j \mathbf{x_j}$.*

Definition 3 ***Minimal Representative Reduced Set*** *(MRRS) is the smallest possible RRS. T_{RRS} is an MRRS of the train set if there exists no other RRS with less cardinality than T_{RRS}.*

Theorem 5.2.1 *Representative Reduced Set (RRS) contains all the support vector indices.*

Proof Let T_{SV} and T_{nSV} be the index sets of support vectors and non-support vector samples, respectively. The direction $\mathbf{w}$ can be written as,

$$w = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i} = \sum_{j \in T_{SV}} \alpha_j y_j \mathbf{x_j} + \sum_{k \in T_{nSV}} \alpha_k y_k \mathbf{x_k} \tag{5.8}$$

Since, $\forall k \in T_{nSV},\ \alpha_k = 0$, $w = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i} = \sum_{j \in U} \alpha_j y_j \mathbf{x_j}$, such that $T_{SV} \subset U$ and $U \subset \{1, 2, \ldots, n\}$.

Therefore, every RRS (set U) contains all the support vector indices, i.e. $T_{SV} \subset T_{RRS}$.

Theorem 5.2.2 *Minimal Representative Reduced Set (MRRS) contains only the support vector indices and maximum cardinality of MRRS is $|T_{SV}|$.*

Proof From Theorem 1, $T_{SV} \subset T_{RRS}$.

The reduced representative set T_{RRS} can further be written as $T_{RRS} = T_{SV} \cup M$, where M contains only the non-support vector indices, i.e. $M \subset T_{nSV}$.

Since the non-support vectors have no impact on the value of $\mathbf{w}$, all of them can be discarded to

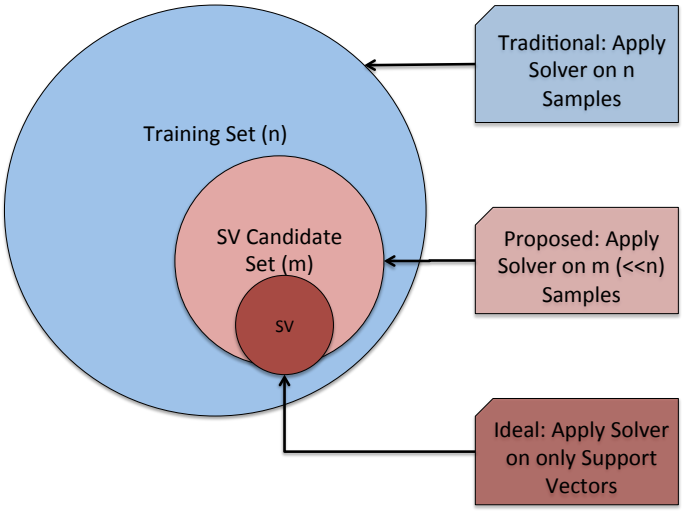


Figure 5-2: Abstract illustration explaining the core concept of the proposed approach, Subclass Reduced Set SVM. Approaches, such as SRS-SVM, that fall under the categorizations of the subset based and piece-wise linear approaches, operate on this basic intuition.

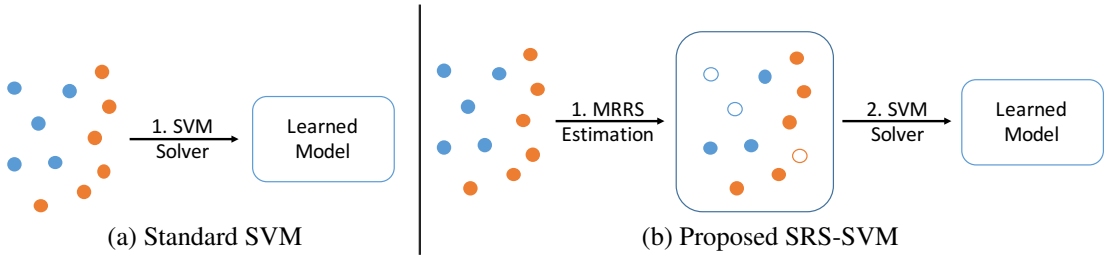
reduce the cardinality of T_{RRS}.

Therefore, if a T_{RRS} is an MRRS, at most, it can contain all the support vector indices and no other indices; i.e. $|T_{RRS}| \leq |T_{SV}|$.

Based on these definitions and Theorems, it can be inferred that 1) RRS would contain *all* the support vector indices and 2) MRRS would contain only the support vector indices. Further, if we can find the RRS, an equivalent decision boundary can be obtained with a relatively smaller set. The MRRS is a smallest such set with which an equivalent decision boundary can be obtained. In the proposed approach, we focus on obtaining the best possible estimate of MRRS in order to reduce the computational time without affecting the classifier performance.

5.3 Proposed Subclass Reduced Set SVM

Theorem 1 implies that if we can estimate the candidate SV set, it can be utilized to obtain the same decision boundary as obtained from the whole train set. If the estimated candidate set contains m samples and $m \ll n$, then the optimization function can be solved with reduced computation and space requirements. In other words, the training time can be reduced significantly, as (1) the number of support vectors is typically very small compared to the total number of training



(a) Standard SVM (b) Proposed SRS-SVM

Figure 5-3: Traditionally SVM solver is applied on the complete training set. The proposed SRS-SVM operates in two stages: estimating MRRS and applying SVM solver on the obtained reduced set. For a detailed illustration of MRRS estimation block, refer Figure 5-4.

samples, i.e. ($|T_{SV}| \ll n$) and (2) the SVM solvers, typically, have quadratic time complexity. Further, leveraging this property is well suited in large datasets, as the inequality $|T_{SV}| \ll n$ is held strongly in densely sampled datasets. Based on this premise, we propose an approach, termed as Subclass Reduced Set SVM (SRS-SVM), to learn SVM with lower training complexity compared to a traditional solver. As illustrated in Figures 5-2 and 5-3, the proposed SRS-SVM has two stages: (1) estimating the MRRS ($|T_{MRRS}| \ll n$) and (2) solving the optimization function on the estimated MRRS. Stage-2 requires less training time as opposed to solving the optimization function on the whole training set; however, a significant training time improvement is achievable only if MRRS is estimated efficiently in Stage-1. Therefore, the proposed approach relies on the efficient estimation of MRRS in order to reduce the overall computational cost.

5.3.1 Estimating Minimal Representative Reduced Set

The detailed concept of the proposed subclass reduced set SVM is illustrated in Figure 5-4. We use piece-wise linearity of nonlinear solutions and the subclass structure of data for estimating MRRS. Details of the MRRS estimation approach are explained below.

- **Leveraging subclass structure of data:** It is well understood that real-world data may form subclasses within a class [172]. Samples sharing some common property may create a subclass within a class. Since, the variation between subclasses is smaller than the variation between classes, subclasses may provide a fine-grained information of the data distribution within a class. Let us consider an example of Dog vs Cat classification problem. There are certain ways in which dogs differ from cats, however, there are certain ways in which one

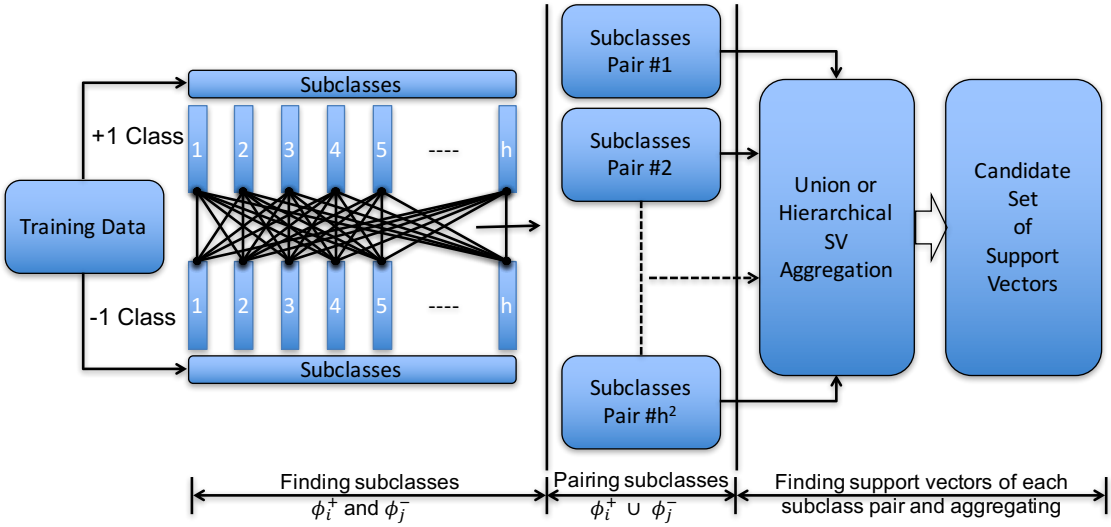


Figure 5-4: Block diagram of MRRS estimation procedure of the proposed Subclass Reduced Set SVM. Each class is divided into h subclasses. Each subclass of $+1$ class is paired with each subclass of -1 class, thus resulting in a total of h^2 subclass-pairs. Support vectors from each subclass-pair are retained as the candidate global support vectors. They are combined either using union operator (in SRS-SVM) or using a further hierarchical aggregation (in Hierarchical SRS-SVM).

breed of dog (e.g. German Shepherd) would differ from another breed of dog (e.g. Doberman Pinscher). In this example, dogs and cats represent classes whereas various breeds represent the subclasses. Figure 5-5 provides further illustration of the applicability of subclass structure for modeling decision boundary. Researchers have attempted to exploit the notion of subclasses for different classifiers [172], [252]. In this research, we explore the subclass notion for fast estimation of MRRS.

As illustrated in Figure 5-5 subclasses represent a finer categorization of a class based on some shared characteristics (in this case, breed of dog). However, the subclass labels are typically not available, therefore, we have to estimate the (pseudo) subclass labels. As each subclass encompasses samples sharing some characteristics, naturally, its estimation is a clustering problem. Subclasses can be obtained with existing approaches such as k-means. Although more sophisticated approaches such as Gaussian mixture modeling may be applied, we have observed that as far as the number of subclasses h is large enough, k-means efficiently estimates MRRS in the proposed framework. Since k-means is an iterative ap-

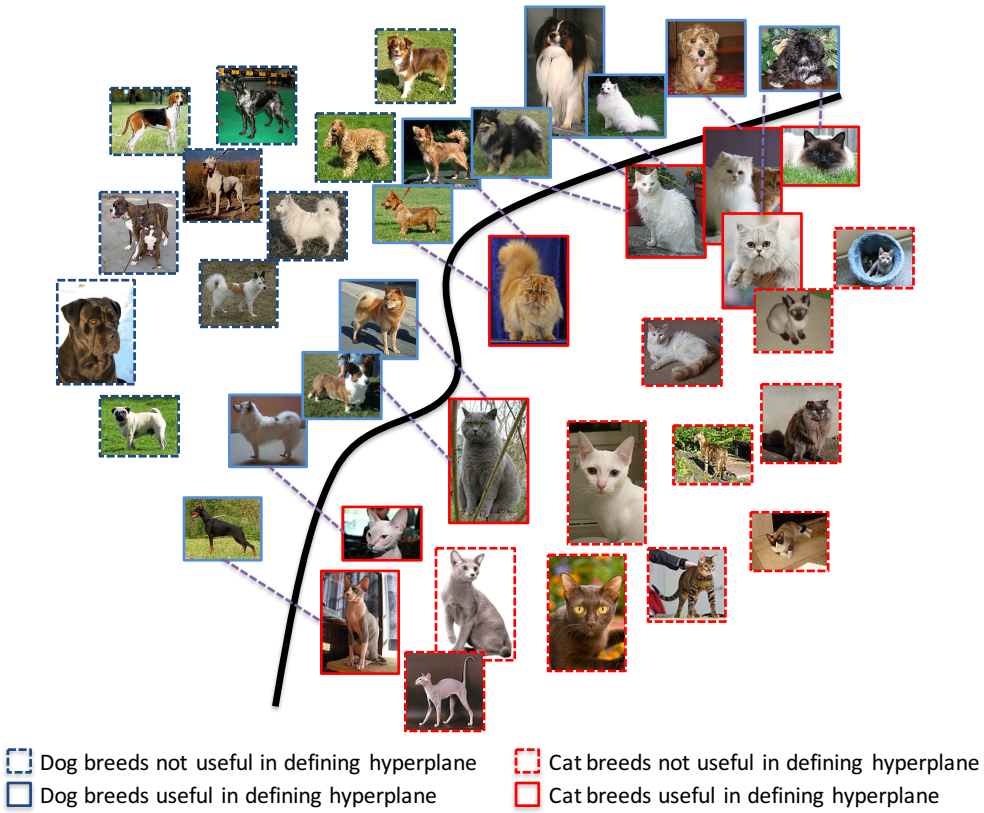


Figure 5-5: Illustrating the applicability of subclass structure in modeling decision boundary for Dog vs Cat classification problem. Out of a vast variety of dog and cat breeds, there are only limited breeds (subclasses) that contribute to the decision boundary. Further, different decision boundaries between breed-pairs of dogs and cats can be seen as constituents for the overall decision boundary.

proach, by restricting the maximum number of iterations, the subclasses can be obtained in relatively less time. By utilizing the Lloyds algorithm [274] the subclasses can be obtained with $O(n^+dhp)$ and $O(n^-dhp)$ time complexities for class $+1$ and -1, respectively. n^+, n^-, d, h, and p represent the number of samples in $+1$ class, the number of samples in -1 class, feature dimensionality, the number of subclasses, and the number of iterations, respectively.

- **Piece-wise linear solution to a nonlinear problem:** It has been suggested in the literature that a nonlinear decision boundary can be achieved with the help of several piece-wise linear solutions (PWL) [253]. This notion also suggests that every piece-wise solution encodes discriminative characteristics of a slice of dataset lying in its vicinity. Essentially, the approach is based on the idea that the nonlinear decision boundary can be approximated by linear boundaries in local regions. Since the decision boundaries are described using support vectors, it implies that the SVs obtained for each local region, jointly, can represent the overall nonlinear decision boundary. Thus, the SVs of piece-wise linear solutions can be utilized to estimate the representative reduced set. It is important to accurately define the local regions for obtaining the linear solutions and subclass structure of the data can be leveraged for this purpose. Theorem 5.3.1 shows that local regions defined as the subclass-pair can be useful in obtaining the global nonlinear solutions.

Theorem 5.3.1 *If a sample is a support vector in the global nonlinear solution, it is a support vector in at least one of the subclass pair-wise solutions.*

Proof If a sample x_i is a support vector in the global nonlinear solution, it is within the margin of the solution.

Therefore, the sample x_p is on the boundary (hull) of its class. [275]

Let x_q be its nearest support vector in the opposite class ($y_p \neq y_q$).

Since x_q is also a support vector, it is on the boundary (hull) of its class.

Without loss of generality, we can assume that x_p and x_q belong to i^{th} and j^{th} subclasses, i.e. $p \in \phi_i^+$ and $q \in \phi_j^-$.

Therefore, x_p is on the boundary (hull) of the i^{th} subclass of $+1$ class and x_q is on the boundary

(hull) of the j^{th} subclass of -1 class, and

x_p is a Support Vector in the solution learned for the subset $\phi = \phi_i^+ \bigcup \phi_j^-$

Theorem 5.3.1 brings together the notion of piece-wise linear solutions and the subclass structure of data by providing the basis for utilizing the subclass structure to obtain the PWL solutions for MRRS estimation. Therefore, the proposed MRRS estimation approach relies on piece-wise linear solutions selected based on the subclass structure. The PWL solutions make it possible to obtain pairs of subclasses that can be utilized to obtain support vectors. Let π be an indicator variable such that $\pi(x_i)$ denotes the subclass association of the i^{th} sample. Let both the classes be divided into h subclasses each[2], $\phi_i^+ = \{k|\pi(x_k) = i \ \ \& \ \ y_k = +1\}$ represents the index set of samples of $+1$ class belonging to the i^{th} subclass, and, similarly, $\phi_j^- = \{k|\pi(x_k) = j \ \& \ y_k = -1\}$ represents the index set of samples of -1 class belonging to the j^{th} subclass, where $i, j \in \{1, 2, 3, \ldots, h\}$. Decision boundaries obtained for the pairs $\phi_i^+ \bigcup \phi_j^-$ describe *a set of possible hyperplanes discriminating two classes in local regions.* All the h^2 pairs of subclasses can be utilized for obtaining the global solution. Estimating minimal representative reduced set requires solving h^2 sub-problems defined on subclass-pairs. Further, it can be easily visualized that if the subclasses are defined in small enough region, the decision boundary for a subclass-pair is likely to be linear. A degenerate case for this is when each sample is considered as a subclass where each subclass-pair solver is bound to yield a linear decision boundary. With approximately reliable subclass association, each subclass-pair decision boundary can be assumed to be linear. Overall, estimation of MRRS involves learning h^2 linear solvers and aggregating their SVs. For simplicity, we assume that each subclass of $+1$ and -1 class have $\frac{n^+}{h}$ and $\frac{n^-}{h}$ samples respectively. As a result, at first, h^2 linear SVM models are learned; each of which is learned over $((n^+ + n^-)/h)$ samples. Note that, this also removes the requirement of storing the whole $(n^+ + n^-) \times (n^+ + n^-)$ kernel matrix in the memory which is a bottleneck for large-scale SVM learning.

Under the assumption of representative subclass categorization and appropriate parameterization, the union set of SVs corresponding to subclass-pair solutions is a representative reduced set. Note that the union set may not necessarily be an MRRS as the mechanism does not prevent a

[2]For the ease of mathematics, we assume that both the classes are divided into equal number of the subclasses. However, that is not a constraint of the proposed SRS-SVM.

Algorithm 2 Proposed Subclass Reduced Set SVM

procedure
Input: Data matrix X, number of subclasses h, cost C, and kernel hyperparameters
▷ Find subclass association of each sample
π=findSubclasses(X) ▷ $\pi(x_i)$ denotes the subclass association of i^{th} sample.
$T_{RRS} = \{\}$ ▷ Initialize reduced representative set
for $i = 1$ to h **do**
for $j = 1$ to h **do**
$\phi_i^+ = \{k|\pi(x_k) = i \ \ \& \ \ y_k = +1\}$ ▷ index set of i^{th} subclass samples of $+1$ class
$\phi_j^- = \{k|\pi(x_k) = j \ \ \& \ \ y_k = -1\}$ ▷ Index set of j^{th} subclass samples of -1 class
$\phi = \phi_i^+ \bigcup \phi_j^-$ ▷ Index set for the subproblem
Solve the subproblem:
$\arg\max_\alpha \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}$, s.t. $0 < \alpha_i \leq C, i, j \in \phi$
$T_{RRS} = T_{RRS} \bigcup \{k|\alpha_k > 0\}$
end for
end for
Solve the nonlinear classification problem on the candidate support vector set
$\arg\max_\alpha \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x_i}, \mathbf{x_j})$, s.t. $0 < \alpha_i \leq C, i, j \in T_{RRS}$
Return: Learned SVM model on T_{RRS}
end procedure

global non-support vector from getting introduced into the union set. However, a large portion of non-support vectors is expected to be absent in the reduced set, yielding a considerable reduction in computational resources required in later stages. In the best case scenario, when no global non-support vector is introduced in the union set, the obtained union set is MRRS, resulting in optimally minimal computation time and space requirements. Algorithm 2 outlines the steps involved in the proposed subclass reduced set SVM.

5.3.2 Hierarchical Subclass Reduced Set SVM (HSRS-SVM)

Consider the most degenerate case, where each class is divided into as many subclasses as the number of samples ($n/2$), implying that each sample belongs to an individual subclass. In this case, each subproblem operates on two samples - one from each class. Both the samples are bound to become support vectors, effectively passing all the training samples into the RRS. Although it is a valid RRS, it is not a good approximation of MRRS. This degenerate case represents the worst case scenario, where the obtained candidate set is same as the whole training set. Further, as shown in Table 5.2, any large value ($\sim \frac{n}{2}$) for h is likely to result in similarly unsuitably very

Table 5.2: The effect of the number of subclasses on the size of estimated MRRS.

	Decreasing number of subclasses $\longrightarrow$				
Subclasses (h)	$\frac{n}{2}$	$\frac{n}{2} - \Delta$	$\cdots$	$h^* + \Delta$	h^*
Size of estimated MRRS	n	$\sim n$	$< n$	$\ll n$	$\ll n$

large MRRS set. At the opposite case, consider a scenario where the whole class is considered as one subclass, i.e. $h = 1$. This configuration is also not useful, as it will violate the assumptions regarding the piece-wise linearity defined on local regions. Thus, very large ($h \approx \frac{n}{2}$) as well as very small ($h \approx 1$) number of subclasses are not likely to yield desirable candidate SV set. In summary, both, overestimation and underestimation of h, are likely to yield sub-optimal results, due to large candidate SV set or basic violation of piece-wise linearity assumptions, respectively.

As the number of subclasses h is varied from $n/2$ (maximum number of subclasses) to h^* (optimum number subclasses), the size of estimated MRRS varies between n and a value close to a total number of global support vectors ($\sim |T_S V|$). The optimal h^* depends on the geometric arrangement of the data; e.g. for `XOR` dataset $h^* = 2$ due to the presence of two distinct clusters for each class. However, for real-world high dimensional datasets, it is crucial to find a reasonably balanced estimate of h.

The solution to the problem is either to estimate h^* or to devise a mechanism that can handle arbitrary higher value of h. Estimating h^* essentially reduces down to understanding the distribution of the class, similar to that in a generative modeling. Since, the philosophical foundations of SVM are in discriminative modeling, we avoid the route of estimating h^*. We focus on creating an extended approach that can provide relatively efficient model even with sub-optimal h. The improved extended approach is a hierarchical version of the proposed approach SRS-SVM. It gains robustness to over-estimation of h by filtering out global non-support vectors at multiple levels of hierarchy.

As shown in Figure 5-6, the mechanism of proposed HSRS-SVM can be described in a tree structure. Since the proposed algorithm follows bottom-up approach, our convention considers the leaf nodes at level 1. Each leaf node caters to one subclass-pair solver $\phi_i^+ \bigcup \phi_j^-$, i.e. a linear SVM is learned within each leaf node. Only the support vectors from each individual solvers is moved

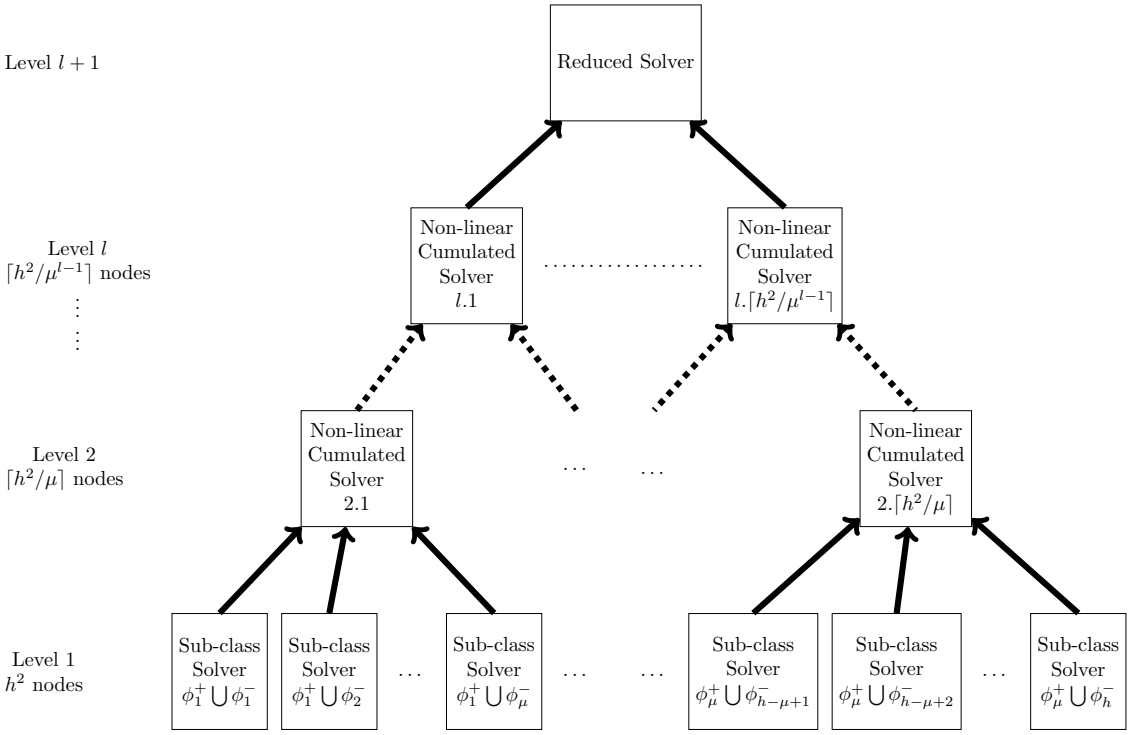


Figure 5-6: Graphical illustration of the proposed Hierarchical Subclass Reduced Set SVM (HSRS-SVM).

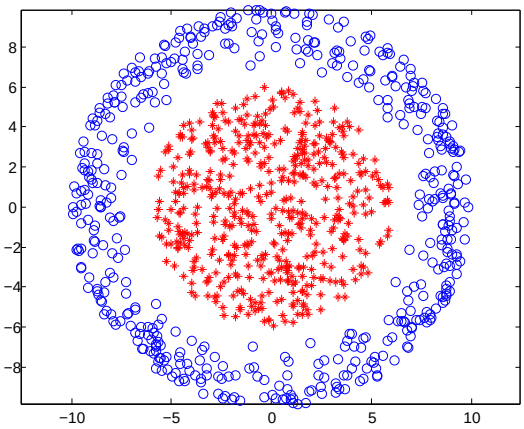
further up in the tree and the remaining samples are discarded. Further, a set of μ models is selected to learn an aggregated solver at the level 2. If each class is divided into h subclasses, there will be h^2 leaf nodes. In this work, a total of $\lceil h^2/\mu \rceil$ aggregated nodes are obtained at level 2. Further, the same aggregation scheme is applied at level 2. Thus, based on the learned $\lceil h^2/\mu \rceil$ models, a total of $\lceil h^2/\mu^2 \rceil$ models are obtained at level 3. In general, the proposed approach operates on $\lceil h^2/\mu^{l-1} \rceil$ nodes at level l. The iterative aggregation stops at the root level consisting of only one node. The model at the root level represents the final aggregated solver model. Since, μ nodes are aggregated at each level, the root node is placed at level k such that $\lceil h^2/\mu^{k-1} \rceil = 1$. Further, in the case of $\mu = h^2$, the root level itself becomes level 2, making the mechanism equivalent to SRS-SVM. Thus, the proposed *SRS-SVM is a special case of HSRS-SVM*.

To increase the chances of introducing samples from various parts of feature space into the next level, nodes are randomly shuffled prior to aggregation. This helps maintain representativeness of the data distribution at next level nodes. For example, without shuffling, the node 2.1 (in Figure 5-6) receives the support vectors from $\phi_1^+ \bigcup \phi_1^-$, $\phi_1^+ \bigcup \phi_2^-$, $\phi_1^+ \bigcup \phi_3^-$, ..., $\phi_1^+ \bigcup \phi_\mu^-$ subclass-pairs. All these subclass-pairs have one common (or repetitively occurring) subclass. The support vectors from these pairs provide a limited view of the overall data spread, as they only encode decision boundary between ϕ_1^+ and the parts of -1 class. Instead, if the nodes are shuffled, a relatively holistic nature of decision boundary may be encoded in the subsequent layers.

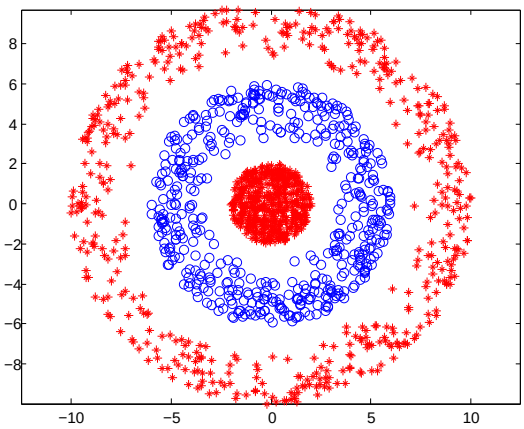
We can learn all the leaf nodes in parallel, as each node corresponds to training a separate linear SVM model. Thus, the total time for leaf level computation is, in the best case scenario, equal to the maximum time required for an individual solver. Further, the level 2 nodes can also be learned in parallel in a similar way. Thus, the total time required for the overall computation is $\sum_{i=1}^{l+1} \max(t_i^1, t_i^2, \ldots)$, where t_i^j is the time required for training j^{th} node in i^{th} level. In practice, propagating the SVs upwards in the tree will also consume computational cycles; however, it will be negligible relative to learning SVM models in each node.

5.4 Datasets and Protocols

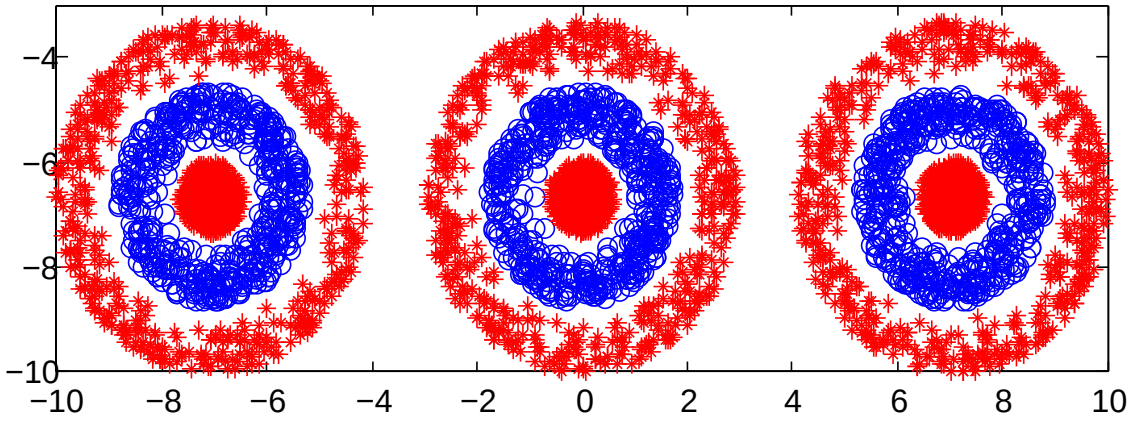
The effectiveness of the proposed SRS-SVM and HSRS-SVM is evaluated on both the non-linearly separable synthetic datasets and real-world datasets. Datasets are chosen with considerable vari-



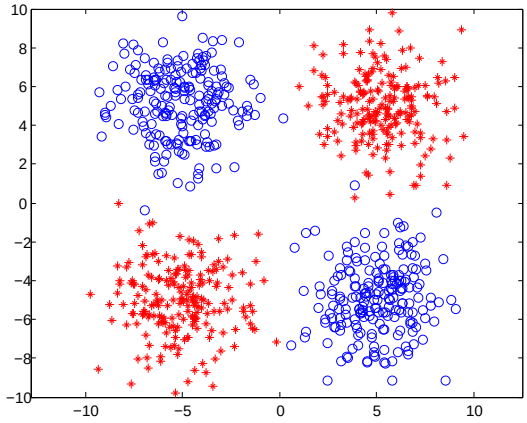
(a) Two Concentric Circles (2CC) ($n = 1000$)



(b) Three Concentric Circles (3CC) ($n = 1500$)



(c) Shooting Range (SR) ($n = 4500$)



(d) XOR ($n = 800$)

Figure 5-7: Illustration the synthetic datasets used for performance evaluation (best viewed in color).

ations in characteristics such as feature dimensionality, training set size, and application domain (finance, weather, object images, face images, textual data) to show the applicability and efficacy of the proposed algorithm.

1. **Nonlinearly Separable Synthetic Datasets:** The synthetic datasets enable performance evaluation in presence of known nonlinearity characteristics. All the synthetic datasets are chosen to be two-dimensional, as they provide an opportunity to visualize the data scatter and the decision boundary.

   (a) Two concentric circles (`2CC`)

   (b) Three concentric circles (or bullseye)(`3CC`)

   (c) Shooting range (a set of bullseyes) (`SR`)

   (d) `XOR` dataset

   Figure 5-7 illustrates the distributions of the above mentioned synthetic datasets utilized in this research. All the synthetic datasets are created by defining the distribution functions. Thus, we can arbitrarily sample varying number of instances from these datasets. Further, the datasets have a varying degree of nonlinearity. For example, the nonlinear nature of the databases increases as we proceed from two concentric circles dataset (`2CC`) to three concentric circles dataset (`3CC`) and then to the shooting range (`SR`) dataset.

2. **Real-world datasets:** The proposed HSRS-SVM approach is evaluated on various real-world datasets. The datasets correspond to classification tasks in different fields of data analytics. The dataset characteristics are described in Table 5.3.

   (a) `adult/census income` [269][3]: predicts whether a person's income exceeds $50K based on various demographic features from census data.

   (b) `ijcnn1` [270][4]: consists of time-series of multiple observations from an internal combustion engine, with the goal of predicting normal and misfiring of the engine.

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#a9a
[4] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#ijcnn1

(a) Animal vs Non-Animals

(b) Face vs Non-Face

Figure 5-8: Samples of the real world databases used for performance evaluation: (a) animal and non-animal class images from `CIFAR-10` [243], [271] and (b) face and non-face images from face detection dataset of Pascal Large Scale Learning Challenge [273]

(c) `covertype` [272][5]: consists of cartographic measures of wilderness areas belonging to seven major forest cover classes. In this work, the dataset is converted to a binary class problem with the goal of separating class 2 from the remaining 6 classes (Protocol used in Collobert, Bengio, and Bengio [276]).

[5] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#covertype.binary

Table 5.3: Details pertaining to the real-world datasets used in the evaluation and their corresponding hyperparameters. (d is feature dimensionality, h is number of subclasses, C is misclassification cost, and γ is radial basis function kernel parameter)

Dataset (size)	**number of training samples**	**number of testing samples**	d	**Parameters**		
				h	C	γ
`adult` (45.8 MB)	32,561	16,281	123	15	1	2^{-5}
`ijcnn1` (23.78 MB)	49,990	91,701	22	5	2^5	2
`covertype.binary` (239.36 MB)	464,810	116,202	54	500	4	2^5
`cifar-10.binary` (1.37 GB)	50,000	10,000	3072	30	2	2^{-22}
`LSL-FD` (1.34 GB)	150,000	50,000	900	50	10	1

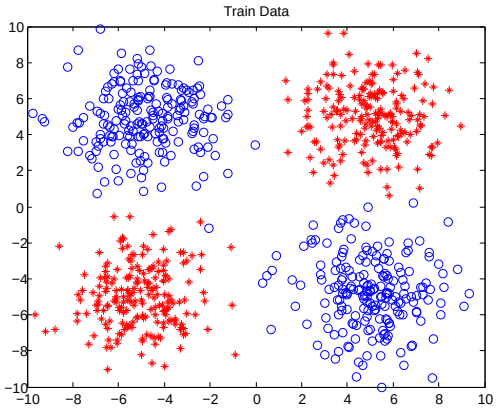
(d) `cifar-10` [271][6]: is an object detection dataset consisting of images of 10 object categories. However, in this work the categories are modified to classify between animals and non-animals (Protocol used in Hsieh, Si, and Dhillon [243]). Figure 5-8(a) shows sample images from both the categories.

(e) Face detection from Pascal Large Scale Learning Challenge (`LSL-FD`) [273]: the dataset consists of a large number of face and non-face images. It is useful for benchmarking face detection performance. Figure 5-8(b) shows sample face and non-face images.

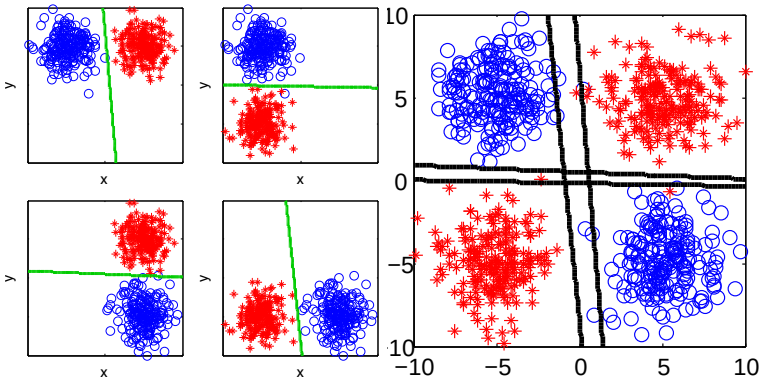
5.5 Experiments on Synthetic Datasets

In the first part of the evaluation, we use synthetic datasets to understand the effectiveness of the proposed approach. As the proposed approach relies on an approximation of original objective functions, the decision boundaries obtained with SRS-SVM are compared with a traditional solver (LibSVM).

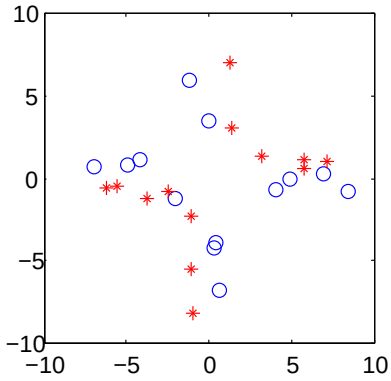
[6] https://www.cs.toronto.edu/~kriz/cifar.html



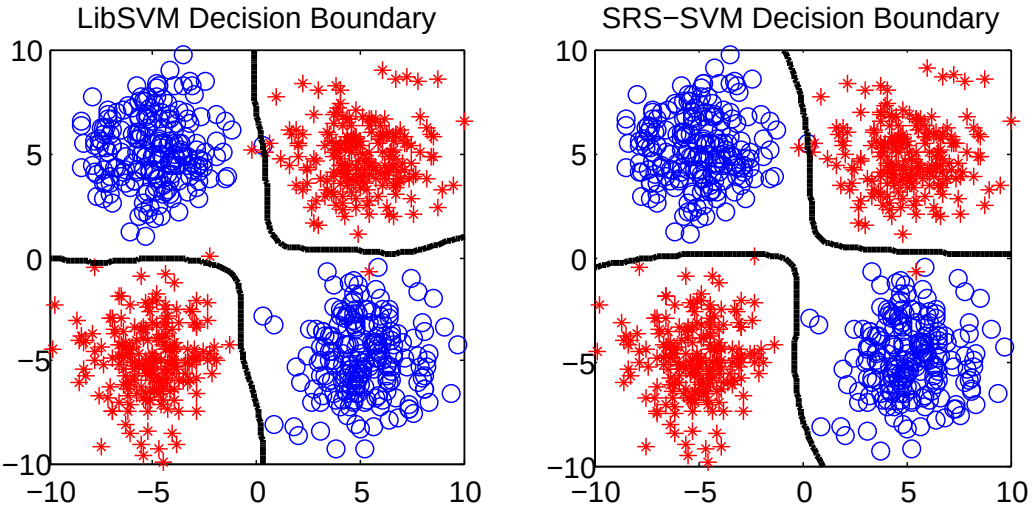
(a) XOR training set



(b) (left) One linear SVM decision boundary is learned for each of the 4 subclass-pairs obtained by dividing each class into 2 subclasses. (right) Set of decision boundaries at Level 1 plotted over test data.

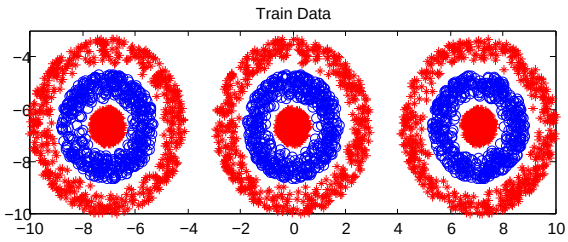


(c) Samples retained at Level 1, i.e. estimated MRRS (Level 1)

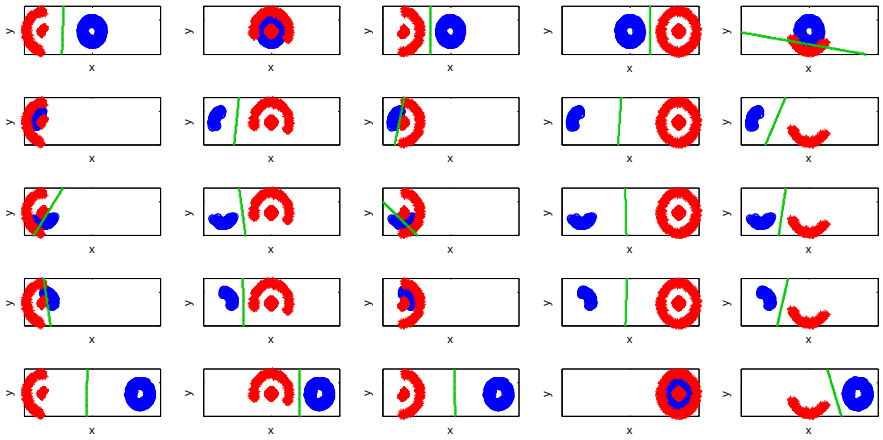


(d) Final decision boundaries (left) obtained with whole data and (right) using proposed SRS-SVM approach.

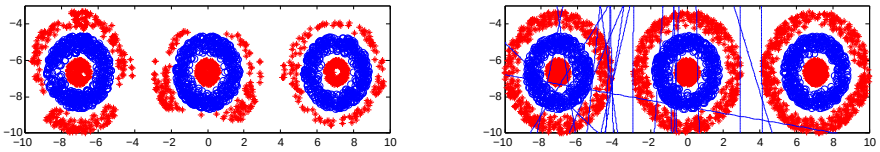
Figure 5-9: Visualization of proposed approach on the XOR dataset. Training on whole dataset ($n = 800, h = 2$) LibSVM takes 3.46 seconds; whereas the proposed SRS-SVM obtains similar decision boundary in 0.25 seconds. See Algorithm 2 to relate the mathematical formulation of the individual steps.



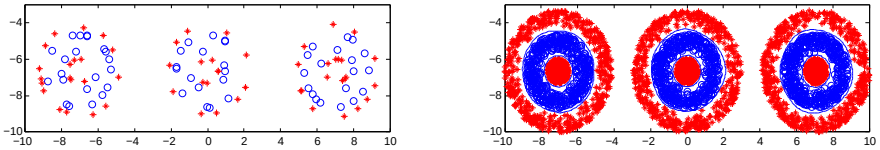
(a) Shooting Range training set



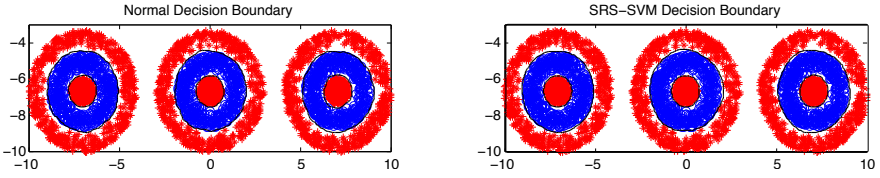
(b) One linear SVM decision boundary is learned for each of the 25 subclass-pairs obtained by dividing each class into 5 subclasses.



(c) (left) Samples retained as candidate SVs at Level 1 (leaf nodes).
(right) Corresponding decision boundaries at Level 1.

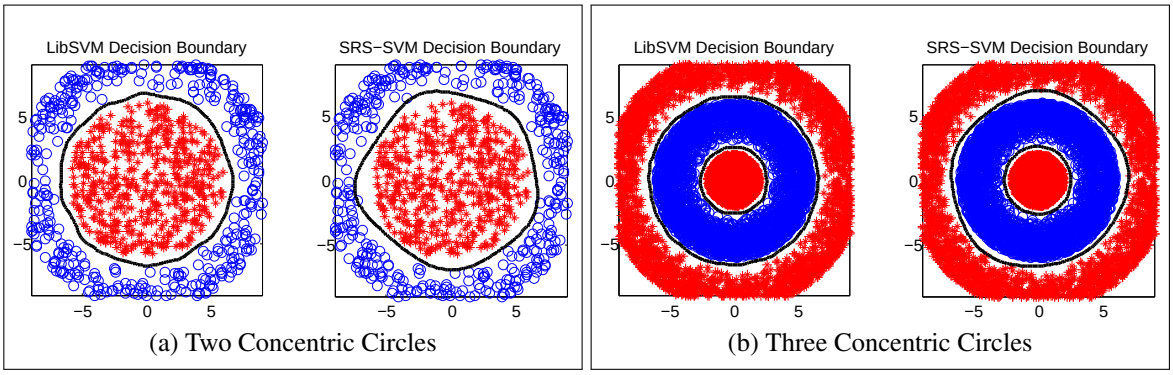


(d) (left) Samples retained as candidate SVs at Level 2 (root node).
(right) Corresponding decision boundary at Level 2.



(e) Final decision boundaries (left) obtained with whole data and
(right) using proposed SRS-SVM.

Figure 5-10: Illustrating the processing of the proposed SRS-SVM on the Shooting Range dataset. Training on the whole dataset ($n = 4{,}500$) LibSVM takes 93 seconds; whereas the proposed SRS-SVM obtains similar decision boundary in 50 seconds.



(a) Two Concentric Circles (b) Three Concentric Circles

Figure 5-11: Comparative illustration of the decision boundaries obtained by LibSVM and by the proposed SRS-SVM approach ($h = 5$).

5.5.1 Visualization of Each Step

We first demonstrate the functioning of the proposed SRS-SVM by providing the visualization of various stages of the algorithm on XOR dataset. The scatter plot of training samples is shown in Figure 5-9(a). The next step involves processing the sub-class pairs with $h = 2$. Figure 5-9(b) shows $h^2 = 4$ subclass-pairs along with a linear SVM decision boundary obtained from each of the subclass-pair based subproblems. All the linear decision boundaries along with the scatter plot of estimated MRRS (candidate SV set) is shown in Figure 5-9(c). Out of $n = 800$ training samples, only 26 are retained as candidate SV set. Thus, a large fraction (96.7%) of samples are discarded at this stage. The final classification boundary obtained using the proposed SRS-SVM is shown in Figure 5-9(d) (right). Comparing this with the decision boundaries obtained by applying LibSVM on the entire training set show that both decision boundaries are very similar for the classification task.

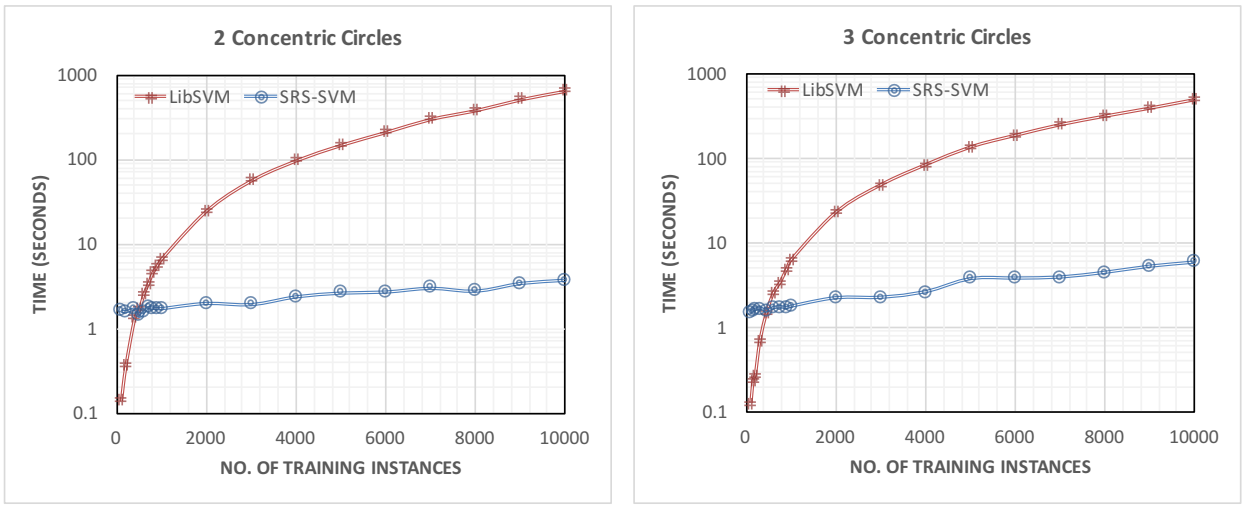
Figure 5-10 shows the working of the proposed SRS-SVM algorithm on the SR (Shooting range) dataset. As the number of subclasses (h) is parametrized to 5, the linear decision boundary is learned for 25 subclass-pairs. It can be observed that a large portion of samples from the outermost band are rejected. The samples lying on the outer boundary of the band are not in the vicinity of the margin of separation, which leads to their rejection as shown in Figure 5-10(c). At the end of Level 1, approximately $3,609$ samples are retained out of the total $4{,}500$ training samples. The SR dataset does not have clearly visible five subclasses; however, due to the mechanism of learning h^2 linear SVMs, the proposed approach yields the decision boundary similar to that obtained with

LibSVM. With minimal reduced representative set (MRRS) estimation, the proposed approach is able to reduce the training time by almost half as compared to LibSVM. Similarly, the decision boundary comparison for the other two synthetic datasets, is shown in Figure 5-11. The XOR dataset actually contains two subclasses, the class corresponding to inner circle of 2CC has actually only one subclass (the class itself), and for 3CC and SR datasets it is hard to concretely define the number of subclasses due to their nonlinearity. However, while applying SRS-SVM, we set the number of subclasses $h = 5$ for all these datasets. Although, it is an inexact parameterization, in all the cases, the decision boundaries obtained with the proposed SRS-SVM are almost same as (visually) those obtained with LibSVM. The efficacy of SRS-SVM with inexact parameterization helps understand its performance in application areas with limited domain knowledge.

5.5.2 Quantitative Analysis

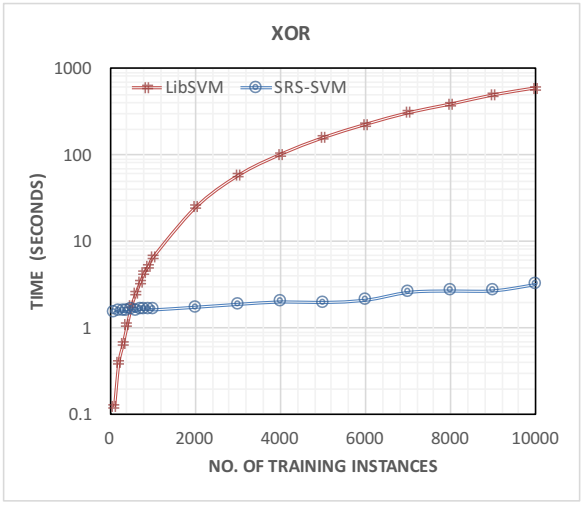
In order to understand the time improvement of the SRS-SVM, we generate varying number of samples from each synthetic dataset. The training time of the proposed approach and LibSVM is compared as a function of the number of training samples. Figure 5-12 shows the graphs corresponding to this experiment for 2CC, 3CC, and XOR datasets. Figure 5-13 shows similar graphs for the SR (shooting range) dataset, with results for additional analysis pertaining to the number of subclass parameter (h).

For all the datasets, both SRS-SVM and LibSVM yield perfect classification on the test sets. The reported training time in this experiment includes the time required for estimating parameters C (misclassification cost) and γ using grid search, and the time required for training the model. It can be observed in Figure 5-12 that for a training size above a certain limit (> 500) the training time of the exact solver (LibSVM) increases rapidly; whereas the rate of increase in the training time is very small in the case of the proposed SRS-SVM. For example, in the case of 2CC dataset with 10,000 samples, the proposed approach requires few seconds ($< 10s$) whereas, the exact solver requires few hundreds of seconds ($< 1{,}000s$) for learning a model. Figure 5-13 shows similar quantitative analysis for Shooting Range dataset. Given that the dataset is relatively complex, we observe that increasing the number of subclasses from 5 to 20, reduces the training time, as it aids in significantly reducing the training set size. For example, for 9,000 training samples, training



(a)

(b)



(c)

Figure 5-12: Comparing training time on three synthetic datasets: two concentric circles (2CC), three concentric circles (3CC), and XOR. A varying number of samples are generated for each of the datasets. The training time is shown on the logarithmic scale. As the number of training instances increases, the training time of LibSVM increases rapidly whereas, the proposed SRS-SVM has a significantly lower rate of increase in training time.

time required for LibSVM is 430.7s; whereas for SRS-SVM with $h = 5, 15, 20$ requires training time of 176.6s, 71.9s, and 64.3s, leading to the speedup of 2.43x, 5.99x, and 6.69x, respectively.

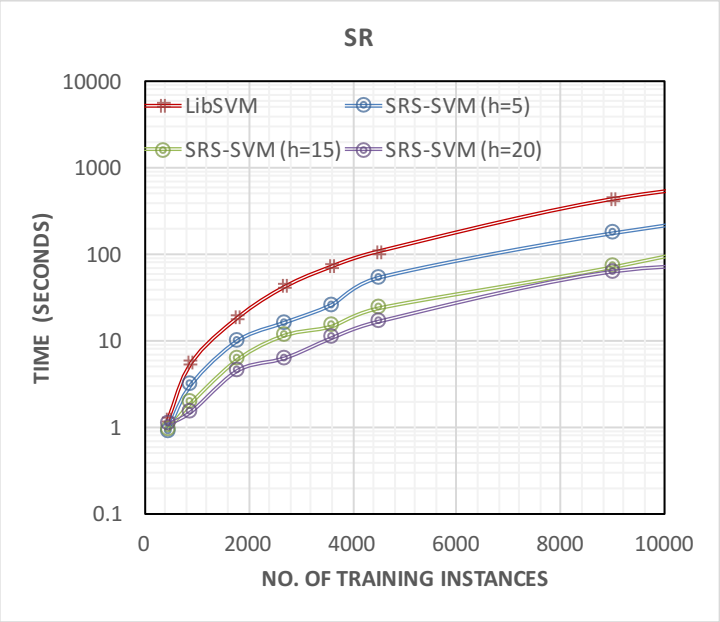


Figure 5-13: Comparing training time on `SR` (Shooting Range) dataset. Different number of samples are generated from the dataset and training set size vs training time plots is shown for different dataset sizes with number of subclasses (h) as 5, 15, and 20. Consistently, SRS-SVM takes less training time compared to LibSVM. As the parameter h is increased, the training time is observed to reduce significantly on the logarithmic scale.

5.6 Experiments on Real-world Datasets

Experiments on diverse real-world datasets are also performed to study (1) the comparative performance of the proposed subclass reduced set based approach, (2) the computational time required at various stages of applying SRS-SVM (namely, clustering, level-1, and level-2), (3) the effectiveness of the proposed representative reduced set (RRS) estimation procedure, and (4) to study the effect of parameters h (number of subclasses) and μ (number of children) on training time and classification accuracy. The first three objectives involve experiments to study the effectiveness of the proposed subclass reduced set based approach with a parameterization of $\mu = h^2$ (and therefore, two levels of hierarchy) as detailed in Section 5.3.1. The experiment is further extended to the proposed hierarchical subclass reduced set SVM (HSRS-SVM) as described in Section 5.3.2.

Table 5.4: Results of the proposed HSRS-SVM in comparison to other related approaches in terms of training time (in seconds) and classification accuracy (in percentage).

(a) Classification Accuracy (%) comparison

Dataset	LibSVM [277]	LLSVM [278]	FastFood [268]	DCSVM [243]	**Proposed** ($\mu = h^2$)
adult	85.01	66.28	85.2	84.75	84.46
ijcnn1	98.70	98.34	91.58	98.39	97.82
covertype.binary	96.07	71.25	out of memory	95.81	93.99
cifar-10.binary	89.66	78.27	79.79	89.78	**89.92**
LSL-FD	99.10	92.27	57.36	99.20	98.50

(b) Training Time (seconds) comparison

Dataset	LibSVM [277]	LLSVM [278]	FastFood [268]	DCSVM [243]	**Proposed** ($\mu = h^2$)
adult	135.4	99.4	83.1	122.6	**60**.2
ijcnn1	68.3	96.6	107.3	74.0	**13**.3
covertype.binary	102,940	1,854	out of memory	75,183	**47,536**
cifar-10.binary	69,128	1,220	459.4	78,107	**38,243**
LSL-FD	311,543	1,396.5	254	515,674	**112,558**

5.6.1 Comparative Analysis

Comparison of the proposed subclass reduced set based approach with existing algorithm is performed with publicly available implementations. [243] have shown that large-scale SVM approaches, namely Cascade SVM [238], SpSVM [267], and core vector machines [235] yield lower accuracies than DCSVM. Therefore, in this work, the results are compared with the most recent approaches namely DCSVM, LLSVM, and FastFood.[7]

1. LibSVM [277]: LibSVM is one of the widely used implementations of SVM that relies on sequential minimal optimization algorithm [269] for optimizing the QP objective function.

2. Divide and Conquer SVM (DCSVM) [243]: DC-SVM is one of the most recent related approaches. In this study, the exact version DC-SVM is utilized.

[7] As the proposed approach relies on accurately finding a subset of the training set, it is logical to investigate the performance of a randomly sampled subset of training set. However, [243] have shown that such random subsets yield suboptimal performance.

3. Low-rank Linearization SVM (LLSVM) [278]: We utilize the LLSVM implementation from the BudgetedSM toolbox [279].

4. FastFood [268]: The technique aims at obtaining approximate high dimensional representation.

Details regarding datasets and the hyper-parameters are provided in Table 5.3. The first set of experiments is performed with parameter $\mu = h^2$, which is a special non-hierarchical case of HSRS-SVM. The results of the comparative prediction performance and training time requirement are reported in Table 5.4(a) and Table 5.4(b), respectively. All the experiments are performed on a Windows machine with two 2.66 GHz Intel Xeon E5640 processors with 48GB primary memory. Compared to LibSVM, the proposed algorithm shows, the speedup of 2.25x, 5.13x, 2.16x, 1.80x, and 2.76x on `adult`, `ijcnn1`, `covertype`, `cifar-10`, and `LSL-FD`, respectively, while yielding similar classification accuracies. Moreover, the speedup of 2.03x, 5.56x, 1.58x, 2.04x, and 4.58x with respect to DCSVM is observed in the case of `adult`, `ijcnn1`, `covertype`, `cifar-10`, and `LSL-FD`, respectively. The basic assumption of the proposed approach is that estimating the candidate support vector set beforehand helps reduce the overall time complexity. The speedup compared to exact solver can be achieved only if the time consumed in estimating the candidate support vector set is lesser than the time saved in learning the SVM model from it. If the dataset is densely sampled, the size of the candidate set is typically a small fraction of the whole training set; almost, guaranteeing improvement in speed. Typically, an exact model learned from a densely sampled set has a relatively very small number of support vectors (e.g. `ijcnn1`, `adult`, and `LSL-FD`) which leads to a significant speed-up with the proposed subclass reduced set based approach.

5.6.2 Training Time of Individual Stage

To further understand the proposed HSRS-SVM approach, we provide its stage-wise training times in Table 5.5. As explained earlier the first stage involves obtaining subclasses, which is followed by Level 1 of training involving the estimation of MRRS based on h^2 linear SVM decision boundaries, and Level 2 involves learning nonlinear decision boundary. Training time of each stage is reported on absolute and relative scale. It is observed that the subclass computation stage takes

Table 5.5: Stage-wise training time of the proposed subclass reduced set based SVM approach. Time is reported in seconds. The figures in the parenthesis represent the fraction of total training time consumed in percentage. Level 2 is the root level as $\mu = h^2$.

Dataset	Subclass computation	Level 1 (MRRS estimation)	Level 2
adult	3.5 (5.8%)	14.8 (24.6%)	41.9 (69.6%)
ijcnn1	1.2 (9.1%)	5.3 (40.2%)	6.7 (50.7%)
covterype.binary	235.9 (0.5%)	2,978.3 (6.3%)	44,315.6 (93.2%)
cifar-10.binary	269.5 (0.7%)	5,855.9 (14.7%)	33,635.4 (84.6%)
LSL-FD	228.7 (0.2%)	42,693.0 (38.1%)	69,636.0 (61.7%)

a very small fraction (0.2-10%) of the total training time. This is a very supportive result as any computationally heavy subclass computation stage can affect the overall computation for large-scale learning. These results also imply that utilizing more time-efficient subclass computation approach may not result in further reducing the training time significantly. Level 1 computation involving MRRS estimation consumes a $6 - 49\%$ of training time. However, this stage involves learning of h^2 linear SVMs independently, thus using parallel architecture (e.g. multi-threading) can further reduce the computation time of Level 1 by multiple folds. Overall, we observe that the Level 2 (i.e. learning nonlinear SVM on estimated MRRS) requires more than 50% of the total training time due to the complex nature of kernel SVM learning.

5.6.3 Effectiveness of MRRS Estimation Approach

This analysis is presented to understand how effectively the proposed subclass based approach estimates the reduced representative set. In order to understand this, its precision and recall are computed with respect to the support vector set (T_{SV}) of the exact solver. If an estimated MRRS ($\hat{T}_{MRRS}$) is a minimal RRS (i.e. smallest possible RRS), it will overlap completely with T_{SV}. Moreover, for an estimated MRRS to have as less spurious candidate support vectors, its *precision*, computed as $\frac{|\hat{T}_{MRRS} \cap T_{SV}|}{|T_{RRS}|}$, should be close to one. Similarly, for an estimated MRRS to have all the actual support vectors, its *recall*, computed as $\frac{|\hat{T}_{MRRS} \cap T_{SV}|}{|T_{SV}|}$, should be close to one.

The precision and recall for the set of SVs in the final SVM model of the proposed approach (T_{rSV}) is also computed. The metrics help in quantifying the similarity between the SVM model of

the exact solver and that obtained with the proposed HSRS-SVM. Note that, this quantification of similarity of two models is independent of the test set. Table 5.6 summarizes the results pertaining to this particular analysis. Key observations are as follows:

- As a general trend it can be observed that recall of estimated MRRS $\hat{T}_{MRRS}$ is high ($> 80\%$) for all the datasets (except `LSL-FD`). This means the proposed MRRS estimation approach retains a large fraction of actual support vectors.

- The basic premise of the MRRS estimation is that it should retain *all* support vectors, i.e. recall is one. The recall of < 1 results from the following two practical aspects: 1) estimating subclasses using a limited iteration approximate k-means without actually modeling the data distribution, and 2) approximating the potentially nonlinear decision boundary of subclass-pairs with a linear decision boundary. Note that both of these approximations yield a significant improvement in training time, with recall > 0.8. Table 5.4 shows that the trade-off does not have a significant impact on the classification accuracy.

- The precision of the MRRS estimation shows that majority of its elements are actual support vectors. A close-to-one precision is not necessary to obtain SVM model equivalent to the traditional solver. However, higher precision of RRS estimate reduces the training time of subsequent levels.

- The precision values of T_{rSV} is typically higher than that of T_{RRS}. This validates the hypothesis that the spurious support vectors in the reduced representative set get discarded in the subsequent levels. Theoretically, the recall of T_{rSV} cannot be higher than that of $\hat{T}_{MRRS}$, as $T_{rSV} \subseteq \hat{T}_{MRRS}$ (therefore, $\frac{|T_{rSV} \cap T_{SV}|}{|T_{SV}|} \leq \frac{|\hat{T}_{MRRS} \cap T_{SV}|}{|T_{SV}|}$).

- In the case of `LSL-FD` dataset, estimated MRRS ($\hat{T}_{MRRS}$) is about half the size of the actual support vector set (T_{SV}). On other datasets, the estimated MRRS is larger than the actual support vector set. Due to this peculiar behavior, we observe that recall values for `LSL-FD` are lower as compared to other datasets. In spite of these observations, the classification performance is affected by only 0.6%, i.e. 99.1% by LibSVM vs 98.5% by the proposed subclass reduced set based approach in Table 5.4.

Table 5.6: Numerical analysis of the precision and recall of the estimated minimal reduced representative set ($\hat{T}_{MRRS}$) and the final support vector set (T_{rSV}) obtained using proposed HSRS-SVM approach with respect to the support vector set (T_{SV}) of the traditional solver (LibSVM).

Dataset	$\lvert T_{SV}\rvert$	$\hat{T}_{MRRS}$ (Estimated MRRS)			T_{rSV}		
		$\lvert\hat{T}_{MRRS}\rvert$	Precision $\frac{\lvert\hat{T}_{MRRS} \cap T_{SV}\rvert}{\lvert\hat{T}_{MRRS}\rvert}$	Recall $\frac{\lvert\hat{T}_{MRRS} \cap T_{SV}\rvert}{\lvert T_{SV}\rvert}$	$\lvert T_{rSV}\rvert$	Precision $\frac{\lvert T_{rSV} \cap T_{SV}\rvert}{\lvert T_{rSV}\rvert}$	Recall $\frac{\lvert T_{rSV} \cap T_{SV}\rvert}{\lvert T_{SV}\rvert}$
`adult`	11,622	13,698	0.7220	0.8509	9,889	0.9093	0.8403
`ijcnn1`	2,478	10,865	0.1913	0.8390	2,202	0.8629	0.8390
`covertype.binary`	98,978	242,998	0.3475	0.8531	88,892	0.8685	0.7800
`cifar-10.binary`	31,750	36,842	0.7197	0.8351	26,616	0.9614	0.8060
`LSL-FD`	130,117	69,991	0.9536	0.5129	67,034	0.9956	0.5129

Table 5.7: Effect of varying number of subclasses (h) and number of children (μ) on the training time and classification accuracy of the proposed HSRS-SVM on the `adult` dataset. The training time is reported in seconds. The figures within parenthesis represent the classification accuracy.

Number of Subclasses (h)	Training Time in seconds (Accuracy in %)					
	$\mu = h^2$	$\mu = \lceil \frac{h^2}{2} \rceil$	$\mu = \lceil \frac{h^2}{4} \rceil$	$\mu = \lceil \frac{h^2}{8} \rceil$	$\mu = \lceil \frac{h^2}{16} \rceil$	$\mu = \lceil \frac{h^2}{32} \rceil$
2	88.0 (68.5)	112.8 (65.8)	n/a			
4	79.0 (82.0)	84.8 (80.7)	84.1 (82.9)	116.0 (77.1)	n/a	
6	68.4 (84.2)	84.2 (83.2)	80.8 (83.5)	103.1 (74.9)	115.8 (75.7)	122.2 (81.7)
8	68.9 (83.7)	98.8 (84.1)	81.2 (83.7)	70.7 (83.9)	94.4 (83.2)	120.2 (78.7)
10	68.1 (83.5)	101.9 (83.2)	82.9 (84.3)	78.2 (82.3)	98.2 (84.2)	116.0 (84.0)
15	70.9 (84.1)	93.2 (84.6)	95.8 (84.3)	80.0 (84.1)	80.2 (84.2)	94.1 (83.5)
20	77.8 (84.7)	111.0 (84.3)	98.0 (84.0)	87.9 (84.7)	82.1 (84.4)	103.0 (84.8)
25	81.4 (84.4)	112.6 (84.7)	106.6 (84.7)	94.7 (84.2)	86.1 (84.7)	112.6 (84.4)
30	88.2 (84.4)	125.0 (84.8)	114.8 (84.9)	107.3 (84.9)	92.4 (84.5)	123.6 (84.5)
35	91.2 (84.7)	132.7 (84.6)	130.0 (84.8)	119.0 (84.7)	102.7 (84.7)	97.7 (84.9)
40	95.5 (84.6)	145.2 (84.8)	139.1 (85.0)	127.6 (84.8)	111.8 (85.0)	100.0 (84.9)
45	102.4 (84.6)	153.1 (84.7)	145.0 (84.6)	138.0 (84.6)	123.7 (84.8)	110.1 (84.7)
50	106.6 (84.9)	162.7 (84.9)	161.0 (84.8)	147.2 (84.8)	133.2 (84.7)	119.4 (84.7)

5.6.4 Effect of h (Number of Subclasses) and μ (Number of Children) Parameters in Hierarchical SRS-SVM

This experiment focuses on understanding the effect of the parameter h (number of subclasses) and μ (number of children) on training time and testing accuracy of the proposed HSRS-SVM. Table 5.2 outlines a theoretical relationship between number of subclasses (h) and size of the estimated MRRS ($|\hat{T}_{MRRS}|$). As explained earlier, a large value of h can render the time improvements

ineffective, whereas a very small value can affect the performance. As detailed in Section 5.3.2, HSRS-SVM can relax the need of fine tuning h by introducing hierarchical structure to the MRRS estimation. The proposed hierarchical structure, which is controlled by μ (number of children), should yield good results with an approximate parameterization of h. This experiment focuses on verifying the expected behavior of the proposed hierarchical SRS-SVM. The number of subclasses h is varied between 2 and 50. For every value of h, experiments are performed with six different values of μ (h^2, $h^2/2$, $h^2/4$, $h^2/8$, $h^2/16$, and $h^2/32$). Since μ has to be a natural number, a ceiling value is used. Table 5.7 summarizes the results for `adult` dataset[8]. However, similar trends were observed on other datasets. Note that, $\mu < h^2/2$ with $h = 2$, and $\mu < h^2/8$ with $h = 4$ are invalid combinations (mentioned as *n/a*) as they do not satisfy the condition $\mu \geq 1$.

- In our experiments, we observe that as the number of subclasses increase, the training time decreases around moderate value ($\sim$15 subclasses) and then increases steadily. The testing accuracy appears to increase rapidly but the rate of increase decreases at higher h approaching saturation. Note that as h increases, so does the size of estimated MRRS which is likely to reduce approximation explaining the accuracy convergence.

- When h is very small, the estimation of MRRS can be poor, i.e. it has low recall (many actual support vectors may be missed) and/or low precision (many non-support vectors are retained). The former will lead to poor testing accuracy, whereas the later will increase the computation time of subsequent levels by increasing the overhead of discarding non-SVs. It can be verified from Table 5.7 that underestimation of h results in overall poor testing accuracy and suboptimal training time.

- Similarly, higher values of h increases the size of estimated MRRS, which affects its precision and overall the training time adversely. However, it improves the recall of MRRS estimation, resulting in the convergence of the decision boundary and testing accuracy to that of an exact solver. As shown in Table 5.7 on `adult` dataset, the classification performance appears to converge/saturate at $h \geq 20$.

- In our experiments, we observe that, for constant h, varying μ from h^2 to $h^2/32$ increases the overall training time because for smaller μ we need to learn more number of intermediate

[8]Due to the exhaustive nature of this experiment, we show tabular results on one dataset.

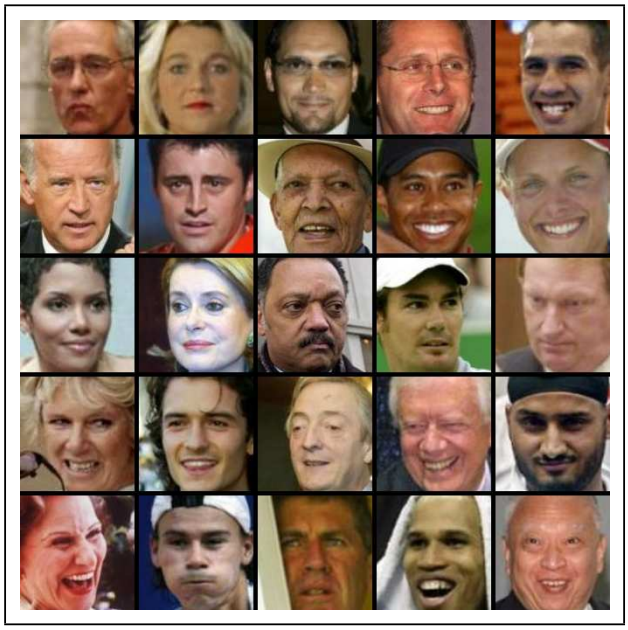


Figure 5-14: Sample images for Labeled Faces in the Wild (LFW) dataset. The classification task involves verifying if the identity of persons in two images is same (match pair) or not (non-match pair).

models. For example, with $\mu = h^2$, h^2 linear SVMs (at Level 1) and 1 nonlinear SVM (at root Level 2) is learned internally; whereas, with $\mu = h^2/2$, h^2 linear SVMs at Level 1, 2 nonlinear SVMs at Level 2, and 1 nonlinear SVM at root Level 3 is computed. This effect is more pronounced with small values of h, as they lead to relatively higher number of samples per subclass; which make the training computationally expensive. However, with higher values of h it is still suitable to set μ at lower values, which can increase prediction performance with relatively less impact on overall training time.

5.6.5 HSRS-SVM with Deep Learning Features for Face Recognition

To further investigate the performance and suitability of the proposed classifier we perform experiments on a challenging problem of face verification. In last few years, deep learning based approaches have established state-of-the-art results in various research areas, especially in computer vision and face recognition. These approaches benefit from utilizing deep learning based features as inputs to traditional classifiers. Therefore, it is our assertion that the proposed subclass reduced set based SVM may also efficiently utilize deep learning based features. Further, this

Table 5.8: Verification accuracy of utilizing LCSSE features with HSRS-SVM ($h = 5, \mu = 25$) and LibSVM in comparison to state-of-the-art approaches.

Approach	Accuracy
LCSSE with HSRS-SVM	**90.92**
LCSSE with LibSVM [165]	90.51
Spartans [280]	87.55
POP-PEP [281]	91.10
MRF-Fusion-CSKDA [282]	95.89

integration of deep learning feature with HSRS-SVM is expected to achieve improved accuracy (by virtue of the features) and to be computationally efficient (by the virtue of the proposed classifier). For face verification, we use Labeled faces in the wild (`LFW`) dataset. Figure 5-14 shows sample images contained in the dataset. The dataset consists of face images with the objective of face verification i.e. predicting match and non-match pairs. The face verification performance is reported for image-restricted protocol. The official protocol defines 10 fold cross-validation splits over 3000 match and 3000 non-match pairs. Each cross-validation contains 5400 images for training and 600 images for testing. We explore the utility of Local Class Sparsity Based Supervised Encoding (LCSSE) [165] which is a deep learning feature representation. The LCSSE feature extractions involves a $l_{2,1}$ norm in auto-encoder based representation learning to promote joint sparsity among same-class samples. Majumdar, Singh, and Vatsa [165] have reported impressive face verification performance using LCSSE features and SVM as classifier. In this experiment, HSRS-SVM is learned over 1,792 dimensional LCSSE feature representations of face images with parameterization of $h = 5$ and $\mu = 25$.

Table 5.8 and Figure 5-15 provides accuracy comparison of LibSVM and HSRS-SVM with same LCSSE feature representations. Further, accuracy values of some of the state-of-the-art approaches are also provided. It is observed that the proposed subclass reduced set based SVM required 2,972 seconds for training whereas, LibSVM required 3,288 seconds. The cardinality of estimated MRRS set is observed to be 4,732. The cardinalities of T_{rSV} (support vectors at root level) and T_{SV} (support vectors of LibSVM) are observed to be 1,809 and 1,874, respectively. It can be seen that the verification performance of proposed HSRS-SVM with the deep learning based features is comparable to state-of-the-art approaches. This provides an empirical evidence

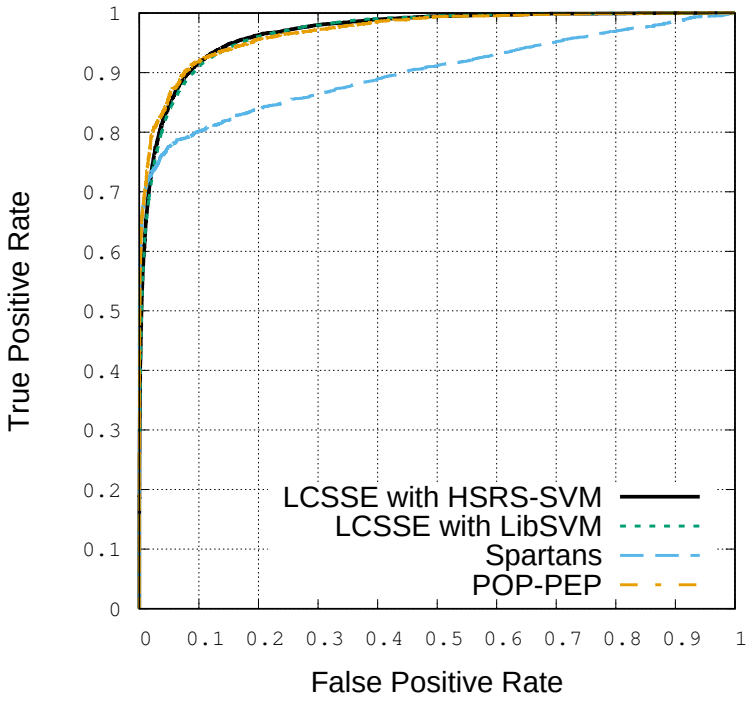


Figure 5-15: ROC curves on restricted protocol of `LFW` dataset.

for the suitability of the proposed approach with deep learning based features.

5.7 Summary

In this work we presented a novel approach for efficiently learning nonlinear support vector machine classifier from large training data. The proposed approach obtains a set of candidate support vectors based on computationally low-cost linear subproblems. We show that utilizing these candidate support vectors (termed as estimated MRRS) to learn the overall nonlinear decision boundary helps to reduce the overall training time significantly. Although, the proposed approach relies on an approximation stage for estimating MRRS, the decision boundary and classification accuracy are not significantly different than that of LibSVM. A hierarchical extension is also proposed, that divides the MRRS estimation task further into multiple iterative stages. Experimental results are shown on several synthetic and real-world datasets including `adult`, `ijcnn1`, `covertype`, `cifar-10`, and `LSL-FD`. Synthetic datasets are leveraged to gain the understanding of individual stages of the proposed approach and to compare the obtained decision boundaries with a traditional

solver. We observe that the proposed approach yields two to five fold speed-up compared to LibSVM and almost up to an order of magnitude compared to other SVM-based large-scale learning approaches. We also showcase the suitability of proposed HSRS-SVM approach with deep learning based features for face verification on `LFW` dataset.

Chapter 6

Conclusion and Future Research Directions

Face recognition has progressed from fascination, to constrained applications, and thence to challenging scenarios such as surveillance. As a culmination of this, researchers are encountering exciting applications as well as arduous challenges associated with large scale applications of unconstrained face recognition. In this direction, this dissertation makes four major contributions: (i) recognize faces with *disguise* variations, (ii) efficiently match identifies with *heterogeneous* (e.g. cross-spectrum, cross-resolution, photo-sketch) face representations, (iii) update the discriminant analysis based face recognition classifiers *incrementally*, and (iv) efficiently learn face recognition classifier from large-scale data.

The first two contributions focus on unconstrained environment where either a user can be uncooperative and uses disguise accessories to hide his/her own identity or the acquisition setting can introduce heterogeneity in the gallery and probe images. To address disguise variations, we propose a novel approach which enhances local region based face classifier with the help of a disguise detection stage. The proposed approach attempts to reject the misleading disguise related facial information and focus only on non-disguised regions for improved face recognition. Experiments are performed on I^2BVSD dataset consisting of 75 subjects. The proposed disguise detection approach achieves up to 85% classification accuracy and the proposed recognition approach outperforms state-of-the-art commercial systems. We have also performed experiments with human annotators which shows that the results of automatic algorithms are similar to unfamiliar face recognition performance of humans. As the second contribution of this research, we present heterogenous discriminant analysis based approach to handle cross-view information such as matching sketches

with digital images and cross-resolution matching in face recognition. HDA and its kernel version (KHDA) encode heterogeneity in the classifier to obtain a common projection space more suitable for matching. Further, we explored the combination of deep learning based feature representation with the proposed HDA/KHDA for heterogeneous face recognition. Experiments are performed on CASIA NIR-VIS-2.0, MultiPIE, and e-PRIP datasets for cross-spectrum, cross-resolution, and photo-to-sketch matching scenarios. On all the three datasets, we report state-of-the-art results; specifically, rank-1 accuracy of 98.1% on the CASIA NIR-VIS 2.0 face database, up to 97.9% on the CMU Multi-PIE database for different resolutions, and 94.7% rank-10 accuracies on the e-PRIP database for digital to composite sketch matching. It would be interesting to explore the proposed approach into other traditional heterogeneous matching scenarios involving pose, illumination, and expression variations.

In the use-case of repeat offenders, the sample images of subjects are available to the face recognition system in the form of incremental batches. A recognition system needs to incrementally update the model based on such incremental data. As the third contribution, we propose Incremental Semi-supervised Discriminant Analysis (ISSDA) approach for face recognition. The traditional subspace learning based approaches rely on updating the between-class and total scatter; on the other hand, the proposed ISSDA utilizes large unlabeled data (in our case a set of unlabeled face images) to estimate the total scatter. The experiments are performed on CMU-PIE, NIR-VIS-2.0, and CMU-MultiPIE datasets. It is observed that ISSDA can update the existing classification model more efficiently as compared to other batch learning and incremental learning subspace approaches. Finally, we propose Subclass Reduced Set Support Vector Machine (SRS-SVM) that can learn from a large-scale training data with less memory and time requirements as compared to traditional solvers. Such a technique allows to efficiently learn face recognition models from very large training sets. The proposed SRS-SVM and its hierarchical extension yield impressive results for various classification tasks and datasets including LFW face dataset. The proposed approach exploits subclass structures of training data to reduce the training set size, which eventually leads to two to ten folds speedup in training time as compared to LibSVM.

With respect to the practical usability of the face recognition systems, there are challenges and covariates that need to be addressed. While this dissertation attempts to address some of the challenges, these solutions may be able to be part of a full fledged system. Such a Universal Face

Recognition (UFR) system would require a set of other solutions too. For example, one of the results we have shown is with respect to cross-resolution matching which assumes the knowledge of the resolution. In general, we make assumptions about *source of heterogeneity*; however, in a full "lights-out" system, this knowledge is typically not available. Similarly, in our work we assume that faces are detected and eye-line aligned. It is possible that in cases such as disguise variations, faces may not be detected under certain disguises. To address such issues it is necessary to have improved face detector for adversarial user behavior and heterogeneity scenarios.

Inspired from the field of big data research, we believe that the next generation face recognition algorithms should encompass 4Vs of face recognition: (1) variety, (2) veracity, (3) volume, and (4) velocity. As shown in Figure 6-1, these 4Vs encompasses the challenges of large scale unconstrained face recognition. Specifically,

- **Variety** refers to the variations in face images, for example, pose, illumination, and ex-

[0] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

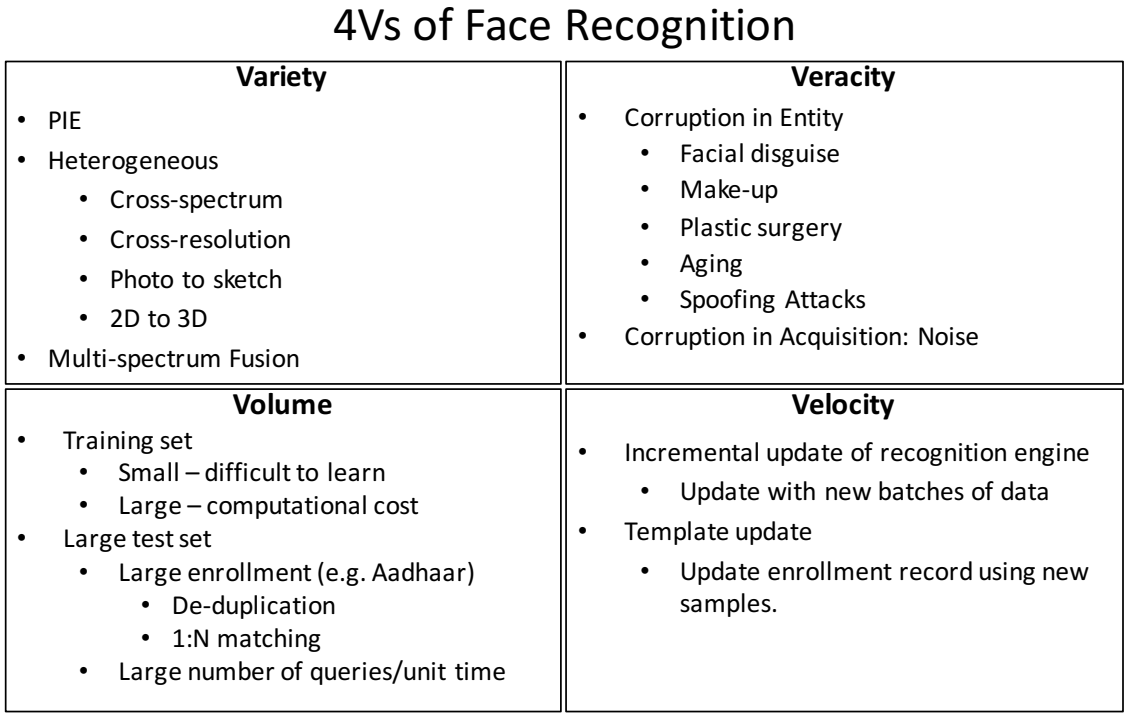


Figure 6-1: 4Vs of face recognition: Classification of face recognition challenges for next generation recognition systems.

pression (PIE). The variations in the image acquisition sensors (visible, near infrared, infrared, 3D, pseudo-3D), and image generation process (hand drawn sketches and composite sketches) also contribute to variety of face images. Other factors contributing to variety of data are distance between camera and face, camera resolution, indoor/outdoor environment, and the time of capture. Two primary approaches to address the *variety* are either to obtain a robust representation or to develop a classifier robust to such variations.

- **Veracity**, in context to face recognition research, refers to abnormalities or extreme corruption of data. The issue of veracity can arise due to various kind of alterations to face sample. Such alterations can be intentional or unintentional and/or reversible or non-reversible. We categorize the problems pertaining to facial disguise, make-up, plastic surgery, aging, and spoofing (print, replay, mask) as prominent veracity challenges. Possible approaches to address these challenges include detecting and discarding the corruption/abnormalities or synthesizing useful information for face recognition.

- **Volume** corresponds to those challenges that are posed by the massive size of the data. The challenges pertaining to large volume, typically, affect the computational time and the space requirements. It can be the volume of the training data or of the query data that may pose challenges for practically usable face recognition systems. Most of the efficient learning algorithms, such as SVM, have super linear time and space complexity with respect to training set sizes and feature dimensionality. Due to this property, most of the learning algorithms scale poorly with massive training sets. Further, the query processing time for identification (1:N matching) and de-duplication scenarios is directly proportional to the enrollment set size. For example, in national identity projects, such as Aadhaar, the de-duplication needs to be performed for the population size of whole nation (approximately 1.2 billion enrolled identities).

- **Velocity** refers to the set of challenges that arises due to the availability of training and/or enrollment data in multiple batches and not in one single batch. Training with multiple batches of data require that 1) the learned model can be updated with the help of new samples and 2) the update requires less time as compared to learning a model on the cumulated training set. In the domain of pattern recognition, this is the basic premise of Incremental

Learning. Another form of velocity related challenge arises with the need to update the existing enrollment samples. For example, aging affects the facial appearance, making it necessary that the face identification record (FIR) is updated at fixed time intervals. This is often referred as Biometric Template Update.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] A. Jain, P. Flynn, and A. A. Ross, *Handbook of biometrics*. Springer Science & Business Media, 2007.

[2] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[3] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar, *Handbook of fingerprint recognition*. Springer Science & Business Media, 2009.

[4] J. Daugman, "How iris recognition works," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 21–30, 2004.

[5] R. B. Hill, "Retina identification," *Biometrics*, pp. 123–141, 2002.

[6] A. Kumar, D. Wong, H. Shen, and A. Jain, "Personal verification using palmprint and hand geometry biometric," in *Audio-and Video-Based Biometric Person Authentication*, Springer, 2003, pp. 1060–1060.

[7] K. Cheng and A. Kumar, "Contactless finger knuckle identification using smartphones," in *International Conference of the Biometrics Special Interest Group*, 2012, pp. 1–6.

[8] 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*, vol. 25, no. 9, pp. 1160–1165, 2003.

[9] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, 2005.

[10] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.

[11] F. Monrose and A. D. Rubin, "Keystroke dynamics as a biometric for authentication," *Future Generation Computer Systems*, vol. 16, no. 4, pp. 351–359, 2000.

[12] A. K. Jain and S. Z. Li, *Handbook of face recognition*. Springer, 2005, vol. 1.

[13] T. Kanade, "Picture processing system by computer complex and recognition of human faces," *Doctoral dissertation, Kyoto University*, vol. 3952, pp. 83–97, 1973.

[14] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[15] P. S. Penev and J. J. Atick, "Local feature analysis: A general statistical theory for object representation," *Network: Computation in neural systems*, vol. 7, no. 3, pp. 477–500, 1996.

[16] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[17] L. Wiskott, J.-M. Fellous, N Kuiger, and C. Von Der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[18] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[19] V. Štruc and N. Pavešić, "Gabor-based kernel partial-least-squares discrimination features for face recognition," *Informatica*, vol. 20, no. 1, pp. 115–138, 2009.

[20] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition*, vol. 33, no. 11, pp. 1771–1782, 2000.

[21] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proceedings of European Conference on Computer Vision*, 2004, pp. 469–481.

[22] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, "On the use of SIFT features for face authentication," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 35–42.

[23] J. Križaj, V. Štruc, and N. Pavešić, "Adaptation of SIFT features for robust face recognition," in *International Conference Image Analysis and Recognition*, Springer, 2010, pp. 394–404.

[24] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2691–2698.

[25] S. Shekhar, V. M. Patel, and R. Chellappa, "Synthesis-based recognition of low resolution faces," in *International Joint Conference on Biometrics*, 2011, pp. 1–6.

[26] V. M. Patel, T. Wu, S. Biswas, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition under variable lighting and pose," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 954–965, 2012.

[27] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[28] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*, 2012, pp. 566–579.

[29] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher Vector Faces in the Wild," in *British Machine Vision Conference*, 2013.

[30] 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*, 2014, pp. 1891–1898.

[31] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *ArXiv preprint arXiv:1502.00873*, 2015.

[32] J. Liu, Y. Deng, and C. Huang, "Targeting ultimate accuracy: Face recognition via deep embedding," *ArXiv preprint arXiv:1506.07310*, 2015.

[33] 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*, 2016, pp. 1–9.

[34] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997.

[35] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[36] 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*, 2015, pp. 815–823.

[37] P. J. Grother, G. W. Quinn, and P. J. Phillips, "Report on the evaluation of 2D still-image face recognition algorithms," *NIST interagency report*, vol. 7709, p. 106, 2010.

[38] P. Grother and M. Ngan, "Face recognition vendor test (FRVT): performance of face identification algorithms," *NIST Interagency report*, vol. 8009, no. 5, 2014.

[39] *Windows hello face authentication*, https://msdn.microsoft.com/en-us/windows/hardware/commercialize/design/device-experiences/windows-hello-face-authentication, [Online; accessed 07-May-2017], 2017.

[40] A. Martinez, "The AR face database," *CVC Technical Report*, vol. 24, 1998.

[41] T. I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa, "Disguise detection and face recognition in visible and thermal spectrums," in *Proceedings of International Conference on Biometrics*, 2013.

[42] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Robust face recognition after plastic surgery using local region analysis," *Image Analysis and Recognition*, pp. 191–200, 2011.

[43] R. Singh, M. Vatsa, H. Bhatt, S. Bharadwaj, A. Noore, and S. Nooreyezdan, "Plastic surgery: A new dimension to face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 441–448, 2010.

[44] J. C. Klontz and A. K. Jain, "A case study on unconstrained facial recognition using the Boston marathon bombings suspects," *Michigan State University, Tech. Rep*, vol. 119, p. 120, 2013.

[45] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "An introduction to the good, the bad, & the ugly face recognition challenge problem," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 346–353.

[46] 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*, 2015, pp. 1931–1939.

[47] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. Duncan, K. Allen, J. Cheney, and P. Grother, "IARPA janus benchmark-B face dataset," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop on Biometrics*, 2017.

[48] 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, Tech. Rep., 2007.

[49] H. S. Bhatt, "Emerging covariates of face recognition," PhD thesis, IIIT-Delhi, 2014.

[50] Y. M. Lui, D. Bolme, B. Draper, J. R. Beveridge, G. Givens, P. J. Phillips, *et al.*, "A meta-analysis of face recognition covariates," in *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2009, pp. 1–8.

[51] *At&t face database*, http://www.cl.cam.ac.uk/Research/DTG/attarchive/facesataglance.html.

[52] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Megaface: A million faces for recognition at scale," *ArXiv preprint arXiv:1505.02108*, 2015.

[53] *Color facial recognition technology (feret)*, http://www.nist.gov/humanid/colorferet.

[54] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[55] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002, pp. 46–51.

[56] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *Audio-and Video-Based Biometric Person Authentication*, Springer, 2003, pp. 44–51.

[57] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE computer society conference on Computer vision and pattern recognition*, vol. 1, 2005, pp. 947–954.

[58] K. Ricanek Jr and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 341–345.

[59] 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," University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007.

[60] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *IEEE International Conference on Computer Vision*, 2009, pp. 365–372.

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

[62] 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*, 2011, pp. 529–534.

[63] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, "Finding celebrities in billions of web images," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 995–1007, 2012.

[64] J. R. Beveridge, J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.

[65] J. R. Beveridge, H. Zhang, P. J. Flynn, Y. Lee, V. E. Liong, J. Lu, M. de Assis Angeloni, T. de Freitas Pereira, H. Li, G. Hua, V. Štruc, J. Križaj, and P. J. Phillips, "The IJCB 2014 PaSC video face and person recognition competition," in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–8.

[66] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *IEEE International Conference on Image Processing*, 2014, pp. 343–347.

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

[68] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, vol. 1, 2015, p. 6.

[69] T. I. Dhamecha, M. Shah, P. Verma, M. Vatsa, and R. Singh, "CrowdFaceDB: Database and benchmarking for face verification in crowd," *Pattern Recognition Letters*, 2017, (Under review).

[70] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[71] B Efraty, C. Huang, S. K. Shah, and I. A. Kakadiaris, "Facial landmark detection in uncontrolled conditions," in *International Joint Conference on Biometrics*, 2011, pp. 1–8.

[72] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2879–2886.

[73] J. R. Beveridge, D. Bolme, B. A. Draper, and M. Teixeira, "The csu face identification evaluation system," *Machine vision and applications*, vol. 16, no. 2, pp. 128–138, 2005.

[74] S. Yoon, J. Feng, and A. Jain, "Altered fingerprints: Analysis and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 451–464, 2012.

[75] G. Righi, J. J. Peissig, and M. J. Tarr, "Recognizing disguised faces," *Visual Cognition*, vol. 20, no. 2, pp. 143–169, 2012.

[76] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: Nineteen results all computer vision researchers should know about," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1948–1962, 2006.

[77] A. Douma, E. Moniz, M. Tarr, and J. Peissig, "Familiarity and the recognition of disguised faces," *Journal of Vision*, vol. 12, no. 9, pp. 980–980, 2012.

[78] E. Moniz, G. Righi, J. J. Peissig, and M. J. Tarr, "The Clark Kent effect: What is the role of familiarity and eyeglasses in recognizing disguised faces?" *Journal of Vision*, vol. 10, no. 7, pp. 615–615, 2010.

[79] U. Toseeb, D. R. Keeble, and E. J. Bryant, "The significance of hair for face recognition," *PloS ONE*, vol. 7, no. 3, e34144, 2012.

[80] H. Leder and C.-C. Carbon, "When context hinders! Learn–test compatibility in face recognition," *The Quarterly Journal of Experimental Psychology Section A*, vol. 58, no. 2, pp. 235–250, 2005.

[81] J. W. Tanaka and M. J. Farah, "Parts and wholes in face recognition," *The Quarterly Journal of Experimental Psychology*, vol. 46, no. 2, pp. 225–245, 1993.

[82] P. J. Hancock, V. Bruce, and A. M. Burton, "Recognition of unfamiliar faces," *Trends in Cognitive Sciences*, vol. 4, no. 9, pp. 330–337, 2000.

[83] S. Dubois, B. Rossion, C. Schiltz, J.-M. Bodart, C. Michel, R. Bruyer, and M. Crommelinck, "Effect of familiarity on the processing of human faces," *Neuroimage*, vol. 9, no. 3, pp. 278–289, 1999.

[84] A. O'toole, K. Deffenbacher, D. Valentin, and H. Abdi, "Structural aspects of face recognition and the other-race effect," *Memory & Cognition*, vol. 22, no. 2, pp. 208–224, 1994.

[85] N. Ramanathan, R. Chellappa, and A. Roy Chowdhury, "Facial similarity across age, disguise, illumination and pose," in *Proceedings of IEEE International Conference on Image Processing*, vol. 3, 2004, pp. 1999–2002.

[86] R. Singh, M. Vatsa, and A. Noore, "Face recognition with disguise and single gallery images," *Image and Vision Computing*, vol. 27, no. 3, pp. 245–257, 2009.

[87] M. De Marsico, M. Nappi, and D. Riccio, "FARO: face recognition against occlusions and expression variations," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 1, pp. 121–132, 2010.

[88] M. Shreve, V. Manohar, D. Goldgof, and S. Sarkar, "Face recognition under camouflage and adverse illumination," in *IEEE International Conference on Biometrics: Theory Applications and Systems*, 2010, pp. 1–6.

[89] A. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, 2002.

[90] P. Belhumeur and D. Kriegman, *The Yale face database*, http://cvc.yale.edu/projects/yalefaces/yalefaces.html, 1997.

[91] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1977–1981, 2005.

[92] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *Proceedings of European Conference on Computer Vision*, 2010, pp. 448–461.

[93] I. Pavlidis and P. Symosek, "The imaging issue in an automatic face/disguise detection system," in *Proceedings of IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, 2000, pp. 15–24.

[94] S. M. Yoon and S.-C. Kee, "Detection of partially occluded face using support vector machines," in *Proceedings of International Conference on Machine Vision Applications*, 2002, pp. 546–549.

[95] J. Kim, Y. Sung, S. Yoon, and B. Park, "A new video surveillance system employing occluded face detection," in *Proceedings of Innovations in Applied Artificial Intelligence*, ser. Lecture Notes in Computer Science, vol. 3533, Springer, 2005, pp. 65–68.

[96] I. Choi and D. Kim, "Facial fraud discrimination using detection and classification," in *Advances in Visual Computing*, ser. Lecture Notes in Computer Science, vol. 6455, Springer, 2010, pp. 199–208.

[97] R. Min, A. Hadid, and J. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition-Workshop*, 2011, pp. 442 –447.

[98] 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*, vol. 39, no. 1, pp. 156–171, 2017.

[99] M. De Marsico, M. Nappi, and M. Tistarelli, *Face Recognition in Adverse Conditions*, ser. Computational Intelligence and Robotics. IGI Global Book, 2014.

[100] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Transactions on Applied Perception*, vol. 8, no. 2, p. 14, 2011.

[101] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Pénard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1642–1646, 2007.

[102] A. J. O'Toole, X. An, J. Dunlop, V. Natu, and P. J. Phillips, "Comparing face recognition algorithms to humans on challenging tasks," *ACM Transactions on Applied Perception*, vol. 9, no. 4, p. 16, 2012.

[103] B. M. 't Hart, T. G. J. Abresch, and W. Einhäuser, "Faces in places: Humans and machines make similar face detection errors," *PloS ONE*, vol. 6, no. 10, e25373, 2011.

[104] A. J. O'Toole, H. Abdi, F. Jiang, and P. J. Phillips, "Fusing face-verification algorithms and humans," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 5, pp. 1149–1155, 2007.

[105] D. Bolme, J. Ross Beveridge, M. Teixeira, and B. Draper, "The CSU face identification evaluation system: Its purpose, features, and structure," in *Proceedings of International Conference on Vision Systems*, 2003, pp. 304–313.

[106] J. Bigun, K.-w. Choy, and H. Olsson, "Evidence on skill differences of women and men concerning face recognition," in *Proceedings of Audio-and Video-Based Biometric Person Authentication*, 2001, pp. 44–50.

[107] F. Gosselin and P. G. Schyns, "Bubbles: A technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, no. 17, pp. 2261–2271, 2001.

[108] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: a robust local image descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705–1720, 2010.

[109] M. De Marsico, M. Nappi, and D. Riccio, "A self-tuning people identification system from split face components," in *Advances in Image and Video Technology*, ser. Lecture Notes in Computer Science, vol. 5414, Springer, 2009.

[110] ——, "Cabala—collaborative architectures based on biometric adaptable layers and activities," *Pattern Recognition*, vol. 45, no. 6, pp. 2348–2362, 2012.

[111] Y. Tajima, K. Ito, T. Aoki, T. Hosoi, S. Nagashima, and K. Kobayashi, "Performance improvement of face recognition algorithms using occluded-region detection," in *Proceedings of International Conference on Biometrics*, 2013.

[112] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[113] P. Grother, G. Quinn, and P. Phillips, "MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms," National Institute of Standards and Technology, NIS-TIR 7709, 2010.

[114] A. M. Burton, R. Jenkins, and S. R. Schweinberger, "Mental representations of familiar faces," *British Journal of Psychology*, vol. 102, no. 4, pp. 943–958, 2011.

[115] C. L. Wilson, P. J. Grother, and R. Chandramouli, "Biometric data specification for personal identity verification," National Institute of Standards & Technology, Tech. Rep. NIST-SP-800-76-1. 2007.

[116] H. Bhatt, R. Singh, and M Vatsa, "Covariates of face recognition," IIIT Delhi, Tech. Rep., 2015.

[117] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Li, "Face matching between near infrared and visible light images," in *Advances in Biometrics*, S.-W. Lee and S. Li, Eds., 2007, pp. 523–530.

[118] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1123–1128.

[119] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1707–1716, 2012.

[120] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1410–1422, 2013.

[121] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 640–652, 2015.

[122] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proceedings of IEEE International Conference on Computer Vision*, 2003, pp. 687–694.

[123] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, "On effectiveness of histogram of oriented gradient features for visible to near infrared face matching," in *IAPR International Conference on Pattern Recognition*, 2014, pp. 1788–1793.

[124] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Advances in Biometrics*, M. Tistarelli and M. Nixon, Eds., vol. 5558, 2009, pp. 209–218.

[125] N. D. Kalka, T. Bourlai, B. Cukic, and L. Hornak, "Cross-spectral face recognition in heterogeneous environments: A case study on matching visible to short-wave infrared imagery," in *IEEE International Joint Conference on Biometrics*, 2011, pp. 1–8.

[126] C. Chen and A. Ross, "Local gradient gabor pattern (LGGP) with applications in face recognition, cross-spectral matching, and soft biometrics," in *SPIE Defense, Security, and Sensing*, 2013, 87120R–87120R.

[127] D. Goswami, C. H. Chan, D. Windridge, and J. Kittler, "Evaluation of face recognition system in heterogeneous environments (visible vs NIR)," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 2160–2167.

[128] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015, pp. 1–7.

[129] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[130] G. Hu, X. Peng, Y. Yang, T. Hospedales, and J. Verbeek, "Frankenstein: Learning deep face representations using small data," *ArXiv preprint arXiv:1603.06470*, 2016.

[131] X. Liu, L. Song, X. Wu, and T. Tan, "Transferring deep representation for nir-vis heterogeneous face recognition," in *IEEE International Conference on Biometrics*, 2016, pp. 1–8.

[132] D. Lin and X. Tang, "Inter-modality face recognition," in *European Conference on Computer Vision*, 2006, pp. 13–26.

[133] S. Z. Li, Z. Lei, and M. Ao, "The HFB face database for heterogeneous face biometrics research," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2009, pp. 1–8.

[134] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 20–23, 2010.

[135] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2011, pp. 593–600.

[136] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[137] Z. Lei, C. Zhou, D. Yi, A. K. Jain, and S. Z. Li, "An improved coupled spectral regression for heterogeneous face recognition," in *IEEE/IAPR International Conference on Biometrics*, 2012, pp. 7–12.

[138] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, "Pose-robust recognition of low-resolution face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 3037–3049, 2013.

[139] S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2019–2030, 2012.

[140] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[141] A. Sharma, A. Kumar, H Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2160–2167.

[142] S. Siena, V. Boddeti, and B. Kumar, "Maximum-margin coupled mappings for cross-domain matching," in *IEEE Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.

[143] Z. Li, D. Gong, Y. Qiao, and D. Tao, "Common feature discriminant analysis for matching infrared face images to optical face images," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2436–2445, 2014.

[144] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Li, "Matching NIR face to VIS face using transduction," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 501–514, 2014.

[145] X. Zou, J. Kittler, and K. Messer, "Ambient illumination variation removal by active near-IR imaging," in *Advances in Biometrics*, Springer, 2005, pp. 19–25.

[146] S. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353.

[147] J. Lu, V. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2041–2056, 2015.

[148] F. Juefei-Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 141–150.

[149] Y. Jin, J. Cao, Y. Wang, and R. Zhi, "Ensemble based extreme learning machine for cross-modality face matching," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11 831–11 846, 2016.

[150] H. Shi, X. Wang, D. Yi, Z. Lei, X. Zhu, and S. Z. Li, "Cross-modality face recognition via heterogeneous joint bayesian," *IEEE Signal Processing Letters*, 2017.

[151] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 301–312, 2017.

[152] P. H. Hennings-Yeomans, S. Baker, and B. V. K. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[153] X. Xing, K. Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognition*, vol. 50, pp. 107–117, 2016.

[154] Y.-H. Yuan, Q.-S. Sun, and H.-W. Ge, "Fractional-order embedding canonical correlation analysis and its applications to multi-view dimensionality reduction and recognition," *Pattern Recognition*, vol. 47, no. 3, pp. 1411–1424, 2014.

[155] X. Chen, S. Chen, H. Xue, and X. Zhou, "A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data," *Pattern Recognition*, vol. 45, no. 5, pp. 2005–2018, 2012.

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

[157] D. Kang, H. Han, A. K. Jain, and S.-W. Lee, "Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching," *Pattern Recognition*, vol. 47, no. 12, pp. 3750 –3766, 2014.

[158] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[159] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational Learning Theory*, Springer, 2001, pp. 416–426.

[160] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[161] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 191–204, 2013.

[162] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa, "Recognizing composite sketches with digital face images via SSD dictionary," in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–6.

[163] B. Klare and A. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *Proceedings of IEEE International Conference on Pattern Recognition*, 2010, pp. 1513–1516.

[164] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[165] A. Majumdar, R. Singh, and M. Vatsa, "Face recognition via class sparsity based supervised encoding," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1273 –1280, 2017.

[166] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, *et al.*, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[167] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image and Vision Computing*, vol. 27, no. 5, pp. 545–559, 2009.

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

[169] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[170] 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*, vol. 32, no. 9, pp. 1627–1645, 2010.

[171] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[172] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1274–1286, 2006.

[173] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 289–302, 2014.

[174] S. Saxena and J. Verbeek, "Heterogeneous face recognition with CNNs," in *European Conference on Computer Vision Workshops*, 2016.

[175] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding," *ArXiv preprint arXiv:1611.06638*, 2016.

[176] Y. Jin, J. Li, C. Lang, and Q. Ruan, "Multi-task clustering elm for vis-nir cross-modal feature learning," *Multidimensional Systems and Signal Processing*, pp. 1–16, 2016.

[177] H. S. Bhatt, R. Singh, M. Vatsa, and N. Ratha, "Matching cross-resolution face images using co-transfer learning," in *IEEE International Conference on Image Processing*, 2012, pp. 1453–1456.

[178] H. S. Bhatt, R. Singh, M. Vatsa, and N. Ratha, "Improving cross-resolution face matching using ensemble-based co-transfer learning," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5654–5669, 2014.

[179] H. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized MCWLD for matching sketches with digital face images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1522–1535, 2012.

[180] B. Klare, Z. Li, and A. Jain, "Matching forensic sketches to mugshot photos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 639–646, 2011.

[181] P. Mittal, A. Jain, R. Singh, and M. Vatsa, "Boosting local descriptors for matching composite and digital face images.," in *IEEE International Conference on Image Processing*, 2013, pp. 2797–2801.

[182] P. Mittal, M. Vatsa, and R. Singh, "Composite sketch recognition via deep network - a transfer learning approach," in *IEEE/IAPR International Conference on Biometrics*, 2015, pp. 251–256.

[183] P. Mittal, A. Jain, G. Goswami, M. Vatsa, and R. Singh, "Composite sketch recognition using saliency and attribute feedback," *Information Fusion*, vol. 33, pp. 86–99, 2017.

[184] T. Chugh, H. S. Bhatt, R. Singh, and M. Vatsa, "Matching age separated composite sketches and digital face images," in *IEEE Biometrics: Theory, Applications and Systems*, 2013, pp. 1–6.

[185] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The FaceSketchID system: Matching facial composites to mugshots," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2248–2263, 2014.

[186] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[187] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.

[188] K. Fukunaga, *Introduction to statistical pattern recognition*. Elsevier, 1990.

[189] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3608–3614, 2006.

[190] H.-T. Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 846–853.

[191] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[192] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers, "Fisher discriminant analysis with kernels," in *IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing*, 1999, pp. 41–48.

[193] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 157–165, 2006.

[194] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji, "Discriminant sparse neighborhood preserving embedding for face recognition," *Pattern Recognition*, vol. 45, no. 8, pp. 2884–2893, 2012.

[195] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–7.

[196] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese, "Semi-supervised local fisher discriminant analysis for dimensionality reduction," *Machine Learning*, vol. 78, no. 1-2, pp. 35–61, 2010.

[197] Y. Zhang and D.-Y. Yeung, "Semisupervised generalized discriminant analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 8, pp. 1207–1217, 2011.

[198] B. Byun, "On discriminative semi-supervised incremental learning with a multi-view perspective for image concept modeling," PhD thesis, Georgia Institute of Technology, 2012.

[199] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–130, 2009.

[200] X. Wu, Y. Jia, and W. Liang, "Incremental discriminant-analysis of canonical correlations for action recognition," *Pattern Recognition*, vol. 43, no. 12, pp. 4190–4197, 2010.

[201] J. Gui, S.-L. Wang, and Y.-K. Lei, "Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data," *Artificial intelligence in medicine*, vol. 50, no. 3, pp. 181–191, 2010.

[202] P. Hall, D. Marshall, and R. Martin, "Merging and splitting eigenspace models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 1042–1049, 2000.

[203] S. Pang, S. Ozawa, and N. Kasabov, "Incremental linear discriminant analysis for classification of data streams," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 5, pp. 905–914, 2005.

[204] H. Zhao and P. C. Yuen, "Incremental linear discriminant analysis for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 1, pp. 210–221, 2008.

[205] P. Howland and H. Park, "Generalizing discriminant analysis using the generalized singular value decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 995–1006, 2004.

[206] L.-P. Liu, Y. Jiang, and Z.-H. Zhou, "Least square incremental linear discriminant analysis," in *IEEE International Conference on Data Mining*, 2009, pp. 298–306.

[207] T.-K. Kim, S.-F. Wong, B. Stenger, J. Kittler, and R. Cipolla, "Incremental linear discriminant analysis using sufficient spanning set approximations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[208] T.-K. Kim, B. Stenger, J. Kittler, and R. Cipolla, "Incremental linear discriminant analysis using sufficient spanning sets and its applications," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 216–232, 2011.

[209] G.-F. Lu, J. Zou, and Y. Wang, "Incremental learning of complete linear discriminant analysis for face recognition," *Knowledge-Based Systems*, vol. 31, pp. 19–27, 2012.

[210] ——, "Incremental complete LDA for face recognition," *Pattern Recognition*, vol. 45, no. 7, pp. 2510–2521, 2012.

[211] H. Lamba, T. I. Dhamecha, M. Vatsa, and R. Singh, "Incremental subclass discriminant analysis: A case study in face recognition," in *IEEE International Conference on Image Processing*, 2012, pp. 593–596.

[212] A. Joseph, Y.-M. Jang, S. Ozawa, and M. Lee, "Extension of incremental linear discriminant analysis to online feature extraction under nonstationary environments," in *Neural Information Processing*, ser. Lecture Notes in Computer Science, T. Huang, Z. Zeng, C. Li, and C. Leung, Eds., vol. 7664, Springer, 2012, pp. 640–647.

[213] Y.-R. Yeh and Y.-C. F. Wang, "A rank-one update method for least squares linear discriminant analysis with concept drift," *Pattern Recognition*, vol. 46, no. 5, pp. 1267–1276, 2013.

[214] G.-F. Lu, Z. Jian, and Y. Wang, "Incremental learning from chunk data for IDR/QR," *Image and Vision Computing*, vol. 36, pp. 1–8, 2015.

[215] J. Ye, Q. Li, H. Xiong, H. Park, R. Janardan, and V. Kumar, "IDR/QR : An incremental dimension reduction algorithm via QR decomposition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1208–1222, 2005.

[216] D. Chu, L.-Z. Liao, M.-P. Ng, and X. Wang, "Incremental linear discriminant analysis: A fast algorithm and comparisons," *IEEE Transactions on Neural Networks and Learning Systems,*, 2015.

[217] T. Hastie, R. Tibshirani, and J. J. H. Friedman, *The elements of statistical learning*. Springer, 2001, vol. 1.

[218] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering.," in *Neural Information Processing Systems*, vol. 14, 2001, pp. 585–591.

[219] ——, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[220] X Niyogi, "Locality preserving projections," in *Neural Information Processing Systems*, vol. 16, 2004, p. 153.

[221] I. Jolliffe, *Principal component analysis*. Wiley, 2005.

[222] X. He, "Incremental semi-supervised subspace learning for image retrieval," in *International Conference on Multimedia*, 2004, pp. 2–8.

[223] Y. Song, F. Nie, and C. Zhang, "Semi-supervised sub-manifold discriminant analysis," *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1806–1813, 2008.

[224] B. Geng, D. Tao, C. Xu, L. Yang, and X.-S. Hua, "Ensemble manifold regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1227–1233, 2012.

[225] J. Gui, W. Jia, L. Zhu, S.-L. Wang, and D.-S. Huang, "Locality preserving discriminant projections for face and palmprint recognition," *Neurocomputing*, vol. 73, no. 13, pp. 2696–2707, 2010.

[226] T. Zhou and D. Tao, "Double shrinking sparse dimension reduction," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 244–257, 2013.

[227] L. N. Trefethen and D. Bau III, *Numerical linear algebra*. Society for Industrial and Applied Mathematics, 1997.

[228] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *IEEE CVPR Workshop on Perception Beyond the Visible Spectrum*, 2013, pp. 1–6.

[229] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *IEEE International Conference on Computer Vision*, 2009, pp. 221–228.

[230] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[231] T. Joachims, "Training linear SVMs in linear time," in *ACM International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 217–226.

[232] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, "A dual coordinate descent method for large-scale linear SVM," in *International Conference on Machine Learning*, 2008, pp. 408–415.

[233] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated subgradient solver for SVM," *Mathematical Programming*, vol. 127, no. 1, pp. 3–30, 2011.

[234] G.-X. Yuan, C.-H. Ho, and C.-J. Lin, "Recent advances of large-scale linear classification," *Proceedings of the IEEE*, vol. 100, no. 9, pp. 2584–2603, 2012.

[235] I. W. Tsang, J. T. Kwok, and P.-M. Cheung, "Core vector machines: Fast SVM training on very large data sets," *Journal of Machine Learning Research*, pp. 363–392, 2005.

[236] H. Yu, J. Yang, and J. Han, "Classifying large data sets using SVMs with hierarchical clusters," in *International ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2003, pp. 306–315.

[237] D. Boley and D. Cao, "Training support vector machines using adaptive clustering.," in *SIAM International Conference on Data Mining*, 2004, pp. 126–137.

[238] H. P. Graf, E. Cosatto, L. Bottou, I. Dourdanovic, and V. Vapnik, "Parallel support vector machines: The cascade SVM," in *Neural Information Processing Systems*, 2004, pp. 521–528.

[239] Y.-J. Lee and O. L. Mangasarian, "RSVM: Reduced support vector machines.," in *SIAM International Conference on Data Mining*, vol. 1, 2001, pp. 325–361.

[240] K.-M. Lin and C.-J. Lin, "A study on reduced support vector machines," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1449–1459, 2003.

[241] P. Ilayaraja, N. Neeba, and C. Jawahar, "Efficient implementation of SVM for large class problems," in *International Conference on Pattern Recognition*, 2008, pp. 1–4.

[242] J. Wang, P. Neskovic, and L. N. Cooper, "Training data selection for support vector machines," in *Advances in Natural Computation*, Springer, 2005, pp. 554–564.

[243] C.-J. Hsieh, S. Si, and I. Dhillon, "A divide-and-conquer solver for kernel support vector machines," in *International Conference on Machine Learning*, 2014, pp. 566–574.

[244] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.

[245] N Syed, H Liu, and K Sung, "Incremental learning with support vector machines," in *International Joint Conference on Artificial Intelligence*, 1999.

[246] "Incremental support vector machine learning: A local approach," in *International Conference on Artificial Neural Networks*, 2001, pp. 322–330.

[247] T Poggio and G Cauwenberghs, "Incremental and decremental support vector machine learning," in *Neural Information Processing Systems*, vol. 13, 2001, p. 409.

[248] M Karasuyama and I Takeuchi, "Multiple incremental decremental learning of support vector machines.," *IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1048–1059, 2010.

[249] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[250] L. Bottou and C.-J. Lin, "Support vector machine solvers," *Large scale kernel machines*, pp. 301–320, 2007.

[251] J. Langford, L. Li, and A. Strehl, *Vowpal wabbit online learning project*, 2007.

[252] N. Gkalelis, V. Mezaris, I. Kompatsiaris, and T. Stathaki, "Linear subclass support vector machines," *IEEE Signal Processing Letters*, vol. 19, no. 9, pp. 575–578, 2012.

[253] X. Huang, S. Mehrkanoon, and J. A. Suykens, "Support vector machines with piecewise linear feature mapping," *Neurocomputing*, vol. 117, pp. 118–127, 2013.

[254] M. Fornoni, B. Caputo, and F. Orabona, "Multiclass latent locally linear support vector machines," in *Asian Conference on Machine Learning*, 2013, pp. 229–244.

[255] L. Ladicky and P. Torr, "Locally linear support vector machines," in *International Conference on Machine Learning*, 2011, pp. 985–992.

[256] V. Kecman and J. P. Brooks, "Locally linear support vector machines and other local models," in *International Joint Conference on Neural Networks*, 2010, pp. 1–6.

[257] T. B. Johnson and C. Guestrin, "Unified methods for exploiting piecewise linear structure in convex optimization," in *Neural Information Processing Systems*, 2016, pp. 4754–4762.

[258] E. Y. Chang, K. Zhu, H. Wang, H. Bai, J. Li, Z. Qiu, and H. Cui, "PSVM: Parallelizing support vector machines on distributed computers," in *Neural Information Processing Systems*, 2007.

[259] S. Tyree, J. R. Gardner, K. Q. Weinberger, K. Agrawal, and J. Tran, "Parallel support vector machines in practice," *ArXiv preprint arXiv:1404.1066*, 2014.

[260] L. Zanni, T. Serafini, and G. Zanghirati, "Parallel software for training large scale support vector machines on multiprocessor systems," *Journal of Machine Learning Research*, vol. 7, pp. 1467–1492, 2006.

[261] T.-N. Do and F. Poulet, "Classifying one billion data with a new distributed SVM algorithm.," in *IEEE International Conference on Research, Innovation and Vision for the Future*, 2006, pp. 59–66.

[262] C. Caragea, D. Caragea, and V. Honavar, "Learning support vector machines from distributed data sources," in *National Conference On Aartificial Intelligence*, 2005, p. 1602.

[263] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *Journal of Machine Learning Research*, vol. 11, pp. 1663–1707, 2010.

[264] T.-N. Do and F. Poulet, "Parallel learning of local SVM algorithms for classifying large datasets," *Transactions on Large-Scale Data- and Knowledge Centered Systems*, pp. 67–93, 2017.

[265] W. Guo, N. K. Alham, Y. Liu, M. Li, and M. Qi, "A resource aware mapreduce based parallel SVM for large scale image classifications," *Neural Processing Letters*, vol. 44, no. 1, pp. 161–184, 2016.

[266] V. Sindhwani and S. S. Keerthi, "Large scale semi-supervised linear SVMs," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 477–484.

[267] S. S. Keerthi, O. Chapelle, and D. DeCoste, "Building support vector machines with reduced classifier complexity," *Journal of Machine Learning Research*, vol. 7, pp. 1493–1515, 2006.

[268] Q. Le, T. Sarlós, and A. Smola, "Fastfood-approximating kernel expansions in loglinear time," in *International Conference on Machine Learning*, 2013.

[269] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," *Advances in kernel methods*, pp. 185–208, 1999.

[270] D. Prokhorov, "IJCNN 2001 neural network competition," *Slide presentation in IJCNN*, vol. 1, p. 97, 2001.

[271] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[272] J. A. Blackard and D. J. Dean, "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables," *Computers and Electronics in Agriculture*, vol. 24, no. 3, pp. 131–151, 1999.

[273] S. Sonnenburg, V. Franc, E. Yom-Tov, and M. Sebag, "Pascal large scale learning challenge," *URL http://largescale.ml.tu-berlin.de*, 2008.

[274] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[275] M. E. Mavroforakis and S. Theodoridis, "A geometric approach to support vector machine (SVM) classification," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 671–682, 2006.

[276] R. Collobert, S. Bengio, and Y. Bengio, "A parallel mixture of SVMs for very large scale problems," *Neural computation*, vol. 14, no. 5, pp. 1105–1114, 2002.

[277] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[278] K. Zhang, L. Lan, Z. Wang, and F. Moerchen, "Scaling up kernel SVM on limited resources: A low-rank linearization approach," in *Artificial Intelligence and Statistics Conference*, 2012, pp. 1425–1434.

[279] N. Djuric, L. Lan, S. Vucetic, and Z. Wang, "BudgetedSVM: A toolbox for scalable SVM approximations," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3813–3817, 2013.

[280] F. Juefei-Xu, K. Luu, and M. Savvides, "Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4780–4795, 2015.

[281] H. Li and G. Hua, "Hierarchical-pep model for real-world face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4055–4064.

[282] S. R. Arashloo and J. Kittler, "Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features," *IEEE Trasactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2100–2109, 2014.