

School of Electrical Engineering and Computing
Department of Computing

Face Image Retrieval with Landmark Detection and
Semantic Concepts Extraction

Antoni Liang

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

May 2017

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgement has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Antoni Liang

Date

Abstract

Automatic identity authentication via biometric verification has been used in a large number of applications, particularly on security systems. Biometric refers to a technology where human unique physical/biological "features" are extracted and analyzed to distinguish an individual's identity among a large group of people. Despite the reliable accuracy, obtaining features such as DNA, fingerprints, and iris is challenging because it requires exclusive devices and participants' cooperation to engage with the devices which leads to a high cost and inconvenience to people. However, facial features do not have these limitations since it can be extracted from a photograph without the subjects' knowledge and the vast development of image-acquisition devices such as surveillance camera and mobile phone leads to an easy access to capture a digital image. This is why applications for automatic face detection/landmarking/recognition are widely popular.

Despite the vast development of facial landmarks detection approaches, most of them emphasize on the application of face recognition or facial expressions recognition with only limited amount of landmarks. Such amount is insufficient for describing the geometric features of facial components such as the shape of the eye. Therefore, semantic-based face application is not feasible in this case since they require a large amount of facial landmarks. The aim of this thesis is to enhance the performances of the frontal facial landmarking system in various ways for the practical application on semantic-based face images retrieval. We proposed several novel approaches based on the state-of-the-art pictorial-tree-structure face models.

First, we propose a dense face model the **AR model** via restructuring a face model architecture with higher density of landmarks information. We emphasize on the crucial components such as eyebrows, eyes, nose and mouth. This model improves the detection accuracy and provides better geometric features. Secondly, we propose the **Multi Resolutions (MR)** landmarking models to detect facial landmarks on low resolution faces as small as 30x30 pixels. We achieved this by proposing an adaptive landmarks scheme for selecting proper facial landmark structure and preserving important landmarks on various face scales. The experiments reveal the high performance on high resolution images and stability on low resolution images. Thirdly, we proposed the lightweight **Tree-structured Filter Model (TFM)** to filter false face detections from the Viola Jones face detector. Additionally, we combine the Viola Jones face detector, TFM and MR models in an integrated system for uncontrolled environment where multiple faces might be present in

the same image on various resolutions. The proposed TFM assists reducing the false face detection rate while maintaining satisfactory true detection rate. Fourth, we extend the utility of pictorial-tree-structure models to glasses/spectacles as the **glasses model** to detect its presence and remove it via image reconstruction approaches, the NLCTV inpainting technique and SFDAE Deep Learning model. The glasses presence is considered as one of the facial semantic features. The experiment shows that the proposed **glasses model** is able to achieve significantly high glasses detection rate along with the corresponding landmarks. Furthermore, our proposed glasses removal system improves both facial recognition and verification rate significantly. Lastly, we evaluate the practicality of our proposed models by investigating the problem of **semantic-based face images retrieval**. For such purpose, we first propose a component-based AR model to further improve the performance of the **AR model**. Then, we utilize the automatically retrieved landmarks to define facial features such as the shape of facial components and glasses presence. We derive some benchmark samples, so we are able to apply semantic mapping for each face as semantic features for face images retrieval. Our experiment demonstrates the feasibility of utilizing semantic features for face images retrieval.

Acknowledgments

I would like to express my greatest gratitude to the following people for the great guidance and support during the long journey of my PhD study. This thesis would not be possible without them.

- First and foremost, I owe my deepest gratitude to my supervisor Associate Professor Wanquan Liu and co-supervisor Associate Professor Ling Li for their constant encouragement, guidance, and support through all my PhD study period. I am deeply grateful for their patience and compassion on teaching me how to be independent and reliable not only as a researcher but also as a person in society. This study has been a very valuable learning experience in my life.
- I really appreciate all the advices and suggestions from Senjian An, Patrick Peursum, David Cooper and Mark Upston during the early years of my study.
- I also greatly appreciate the support and care from my parents Fransisca Zainiada Usman and Liong Bie Siong, my brother Johan Liang, and my sister Yenni Liang. Thank you so much for keep motivating and believing in me. I am also grateful to my best friend Anggiria Lestari Megasari for the support.
- I would like to thank all my fellow research students (Master and PhD) and researchers (Nadith Pathirage, Qilin Li, Chenyu Wang, Xin Zhang, Jinglan Tian, Mustafa M. M. Alrjebi, Xiaoming Chen, Xiang Xu, Ming Liang, Ke Fan, Billy Li, Ren Yan, Mary, Jingjing Liu, Ruhua Wang, Xianchao Xiu, Shichu Chen, Mohammed Ahmed Bakhrayba, Lloyd Compelio, and Mir Rizwan Farid) for all the inspirations and support.
- I would also like to thank our administrative officers Mary Mulligan, Patricia Tang, and Esther Yew for your assistance on the paper works.
- Lastly, thank you so much to Curtin University for providing financial support for my PhD study through International Postgraduate Research Scholarship (IPRS).

Published Work

Several academic publications included in this thesis over the course of my PhD study are listed in chapter order:

- Antoni Liang, Wanquan Liu, Ling Li, Mir Rizwan Farid and Vuong Le. (2014) Accurate Facial Landmarks Detection for Frontal Faces With Extended Tree-structured Models. *22nd International Conference on Pattern Recognition (ICPR)*. pages 538-543. (Chapter 3)
- Antoni Liang, Chenyu Wang, Wanquan Liu, and Ling Li. (2014) A Novel Landmark Detector System for Multi Resolution Frontal Faces. *International Conference on Digital Image Computing: Techniques and Applications (DlCTA)*. pages 1-8. (Chapter 4 and 5)
- Antoni Liang, Chenyu Wang, Wanquan Liu, and Ling Li. (2016) Robust and Flexible Landmarks Detection for Uncontrolled Frontal Faces in the Wild. *Numerical Algebra, Control and Optimization (NACO) Volume 6*. pages 263-296. (Chapter 4 and 5)
- Antoni Liang, Chathurdara Sri Nadith Pathirage, Chenyu Wang, Wanquan Liu, Ling Li, and Jinming Duan. (2015) Face Recognition Despite Wearing Glasses. *International Conference on Digital Image Computing: Techniques and Applications (DlCTA)*. pages 1-8. (Chapter 6)
- Antoni Liang, Wanquan Liu and Ling Li. (Under Preparation) Role of Semantic Features on Face Images Retrieval Based on Automatic Facial Landmarks Detection. (Chapter 7)

Contents

1	Introduction	1
1.1	Research Gaps and Aims	2
1.2	Thesis Structure and Contributions	4
2	Background	8
2.1	Pictorial-Tree-Structured Face Models	8
2.1.1	Pictorial Structure	8
2.1.2	Histogram of Oriented Gradients (HOG)	11
2.1.3	Zhu and Ramanan Face Models	14
2.1.4	Source Code and Pre-trained Models	16
2.2	The Viola and Jones Face Detector	18
2.3	Image Reconstruction Approaches	21
2.3.1	The NLCTV (Non-Local Colour Total Variation) Inpainting	21
2.3.2	The SFDAE (Stacked Face De-noising Auto Encoders) Deep Learning Model	24
2.4	Face Recognition Techniques	26
2.4.1	The PCA (Principal Component Analysis)	27
2.4.2	The LDA (Linear Discriminant Analysis)	28
2.4.3	The SRC (Sparse Representation Classifier)	29
2.5	Databases	31
2.5.1	Controlled Face Databases	31
2.5.2	Uncontrolled Face Database	38
2.5.3	INRIAperson	39
2.6	Summary	39
3	Facial Landmarks Detector with High Density Landmarks	42
3.1	Model Creation	45
3.2	Model Training and AR Database	45
3.3	Experiments	47
3.3.1	Evaluation Protocols	47
3.3.2	The Independent-1050 Model VS the AR Model	52
3.3.3	The CompASM model VS AR Model	53
3.3.4	The AR Model with Different Colour Spaces	57
3.4	Summary	57

4	Facial Landmarks Detection for Multi-Resolutions Images	59
4.1	The Multi-Resolutions (MR) models	60
4.1.1	Adaptive Number of Landmarks via Resolution Reduction	61
4.1.2	Training for the MR Models	65
4.2	Experiments	69
4.2.1	Testing Dataset	70
4.2.2	The Evaluation Protocols	70
4.2.3	The MR Models VS the Share-146 Model	70
4.2.4	The MR Models VS Other State-of-the-art Approaches	73
4.3	Summary	78
5	Fast and Effective Face Detector	81
5.1	The Tree-structured Filter Model (TFM)	82
5.1.1	Model Training	85
5.1.2	Experiment Setup	85
5.1.3	Experiment Results	87
5.2	Facial Landmarking System	88
5.2.1	Comparison with the Share-146	89
5.2.2	Speed Comparison	89
5.3	Summary	96
6	Glasses Detection and Removal for Face Recognition and Verification	98
6.1	Framework	100
6.1.1	Face Alignment	100
6.1.2	Glasses Model	102
6.1.3	Masking	103
6.1.4	The Complete Framework	104
6.2	Experiments	106
6.2.1	Glasses Detection/Landmarking	110
6.2.2	Glasses Removal/Reconstruction	111
6.3	Summary	118
7	Face Retrieval based on Semantic Features via Face Landmarking	120
7.1	Framework	121
7.1.1	Face Database	121
7.1.2	Facial Landmarks Extraction	123
7.1.3	Semantic Features	126
7.1.4	Semantic Mapping	135
7.2	Experiments	137
7.2.1	Experiment Setup and Performance Evaluation	137

7.2.2	Experiment Results	138
7.3	Summary	142
8	Conclusions and Future Directions	143
8.1	Future Study	145

List of Figures

1.1	Visual cues of the remaining chapters. Starting from the top moving clockwise, we briefly discuss about the related approaches/techniques and relevant databases used in our experiments in Chapter 2. Our research is mainly focused on pictorial-tree-structured models to detect landmarks on facial region with various implementations. In Chapter 3, we proposed robust frontal face models with higher accuracy and condensed landmarks in order to extract semantic features from the human faces. We further expand the capability of our face models in Chapter 4 to cover various resolutions as low as 30x30. In order to have an efficient facial landmarks detection framework, we propose light-weight face models for fast initial face detection in Chapter 5. In Chapter 6, we propose novel glasses landmarking models as facial glasses can be considered as a part of the face and its presence is highly common. Lastly, we apply our proposed face and glasses models to extract facial semantic features and conduct facial images retrieval in Chapter 7. The conclusion and future directions are summarized in Chapter 8	7
2.1	These are 2D triangle shapes in various orientations, sizes, scales/ratios, and colours. Even with these large varieties of deformation, people still recognize the triangle shape.	9
2.2	A pictorial structure of a standing human. The whole body can be represented by ten parts: head, torso, 2 upper arms, 2 lower arms, 2 upper legs, 2 lower legs.	10
2.3	(a) Original image (b) Rectangular partitions of the image called "cells". .	11
2.4	Pixel intensity differences can be calculated by applying masks on both horizontal (d_x) and vertical (d_y) directions. Both values can be used to compute the magnitude/strength and the orientation/angle (θ) of the gradients. . . .	12
2.5	Examples of blocks with 2x2 cells. Block 1 covers the edge of the roof with 2 distinct orientations which is shown as 2 lines in the corresponding HOG feature. Block 2 only contains 1 orientation.	13
2.6	6 mixtures of tree structure for 6 variations of facial expressions (neutral, smile, surprise, squint, disgust, scream) from CMU multiPIE database (Gross <i>et al.</i> , 2010). (© 2014, IEEE)	14

2.7	Each face model involves 2 main components. The first component is the appearance which describes the HOG features (Dalal and Triggs, 2005) on each part of the tree structure. The second component is the comprehensive arrangement of the relations among all the parts which defines the overall tree shape of the face.	15
2.8	<i>Independent-1050</i> face models proposed by (Zhu and Ramanan, 2012a) with 18 mixtures. It covers 13 facial poses including frontal poses and 5 facial expressions.	17
2.9	Some examples of Haar features used in Viola & Jones face detector.	18
2.10	(a) "Integral image": the value of any position is the sum (\sum) of all pixel image intensities on the left or top of it. (b) Sum of image intensity on any region can be done by basic arithmetical operations on all four corner values. In this example, region D can be computed as $sum = D - C - B + A$	19
2.11	(a) Original face image. (b) Haar feature indicates the presence of eyes by emphasizing the dark region on eyes and bright region on upper cheeks. (c) Haar feature indicates the presence of eyes by emphasizing the bright part on nose bridge covered by dark region on both sides.	20
2.12	The classifier has multiple stages. If the input sub-window fails in any stage, it will be considered a non-face region. This allows for a quick elimination procedure and assigning more computation on the face regions.	20
2.13	These images are the iterative process of NLCTV inpainting approach on a sky image. The image was tainted by irregular dark lines across the whole image (first image). The last image is the original image.	23
2.14	This is the overview of SFDAE framework. f_1 represents the low dimensional features derived from the first layer. The features are processed through de-noising process on second layer represented by f_2 . Lastly, the features are decoded to reconstruct the frontal neutral faces.	25
2.15	13 image variations on 2 sessions for each participant on AR database.	32
2.16	Landmarks ground truth on AR database.	33
2.17	Examples of all the facial poses and expressions on CMU multiPIE database.	34
2.18	Sample images from PUT database. Every 3 images is one subset.	34
2.19	Sample images from FEI database.	35
2.20	3D and 2D data from BU-4D database.	35
2.21	range data (depth) and 2D images from CurtinFaces database.	36
2.22	Images with variety on (from top) pose, accessories, expression, and illumination from CAS-PEAL-R1.	37
2.23	Some collections of face images from FDDB database.	38
2.24	Samples of images from AFLW database.	40
2.25	Samples of non-facial images from INRIAperson database.	41

3.1	Facial Landmarks from the Independent-1050 model is not sufficient to extract semantic features. (© 2014, IEEE)	43
3.2	Examples of the facial landmarking results between Independent-1050 models (top) and our proposed model (bottom). (© 2014, IEEE)	44
3.3	The tree structure of Independent-1050 model (left) and proposed AR model (right). The tree restructuring was made to depict better geometric descriptions and improve the accuracy rate. (© 2014, IEEE)	46
3.4	Visualization of frontal face models with various facial expressions from Independent-1050 (top) and AR model (bottom). (© 2014, IEEE)	48
3.5	3 different thresholds for detection rate on eye corner. Starting from the smallest circles are the 5%, 10%, and 20% of IOD respectively. (© 2014, IEEE)	49
3.6	15 landmarks for comparison on Independent-1050 (left) and 17 landmarks for comparison on CompASM and AR model (right).	50
3.7	Width and height of the facial components are calculated as the largest distance between landmarks on horizontal (x-axis) and vertical (y-axis) directions respectively.	51
3.8	Four colour spaces: RGB, grey-scale, HSV, and RGB-NII.	51
3.9	Some testing result from Independent-1050 (left) and AR model (middle). The landmarks in ground truth are shown in the last column (right).	54
3.10	Samples of CompASM results on scream expression. The facial landmarking performance is not so accurate on scream expression.	56
4.1	A large gap created every time a landmark is removed. (© 2016, AIMS)	62
4.2	18 chosen Very Important Points (VIP) to preserve on the proposed MR models. (© 2014, IEEE. 2016, AIMS)	63
4.3	Face structure is divided into 17 segments by the VIP. (© 2014, IEEE. 2016, AIMS)	64
4.4	Rearrangement of the landmarks when performing landmarks reduction. Red dots represent the initial landmarks while the green dots represent the revised landmarks. (© 2016, AIMS)	65
4.5	The landmarks reduction process on a face. We emphasize on the facial components eye, nose, and mouth in this figure. The order of the scale level is as follows: 100% (ground truth), 70%, 50%, 30%, 10%. The face images were not scaled down here for easier observation. (© 2014, IEEE)	67
4.6	The complete set of MR models. Starting from the first row are the MR-14, MR-36, MR-70, MR-103, and MR-130. Various facial expressions are shown in the order of neutral, smile, angry, and scream. (© 2016, AIMS)	68

4.7	4 VIP are exempted from the MR-14 model. Our observation shows that the features are too subtle to be included. (© 2016, AIMS)	69
4.8	Samples of face image in various resolutions. In clockwise direction, the sizes shown here are 750x750, 600x600, 450x450, 300x300, 210x210, 150x150, 90x90, and 30x30. It is clearly seen that the information difference between large and small faces are imminent. (© 2014, IEEE)	71
4.9	Eleven facial landmarks for performance evaluation. (© 2014, IEEE. 2016, AIMS)	72
4.10	Relative error on PUT database.	74
4.11	STASM is susceptible to detecting facial landmarks incorrectly if it is employed on the non-face regions. Face detection with very high accuracy is required in this case. (© 2016, AIMS)	76
4.12	Detection rate on 5% IOD threshold.	77
4.13	Detection rate on 10% IOD threshold.	77
4.14	Detection rate on 20% IOD threshold.	78
4.15	Examples of facial landmarks misalignments which occur on Intraface (second column) and STASM (third column) with 30x30 faces. Despite slightly less accurate, MR models have a major advantage of robustness against misalignment on facial components especially with the presence of beard and slight occlusion of eyebrows. (© 2016, AIMS)	79
5.1	This is the illustration on how the Viola Jones (VJ) face detector performs together with Tree-structured Filter Model (TFM) concluded with facial landmarking by MR models. In this particular example, VJ successfully detect all 4 faces, but with the expense of 7 false positives. TFM then rapidly examines all the face candidates, successfully removing 6 false detections while maintaining all true detections. The last false positive is then disregarded by MR models. Since TFM has removed most of false detection quickly, it reduces the workload of MR models. (© 2014, IEEE)	83
5.2	Visualization of TFM on neutral (left) and scream (right) facial expressions. (© 2014, IEEE. 2016, AIMS)	86
5.3	Some chosen frontal faces from FDDb database. (© 2016, AIMS)	87
5.4	(Left) An example of query. (Top right) A face detected by Viola Jones and expanded prior to filtering by TFM. The subwindow is expanded to ensure sufficient coverage of the whole face for the facial landmarking phase. (Bottom right) A subwindow of a face detected by Share-146 model. It is cropped based on the edge of landmarks and nose tip as a central part to include forehead region.	88
5.5	ROC (Receiver Operating Characteristic) Curve for Various Face Detectors.	89

5.6	Time Comparison between VJ, VJ+TFM, and SHARE-146.	90
5.7	(Left) The face candidates detected by Viola Jones detector. (Right) False positives are removed by our proposed TFM. (© 2016, AIMS)	91
5.8	The integrated system combined from Viola Jones detector, TFM, and MR models. (© 2016, AIMS)	92
5.9	Chosen images from AFLW database. (© 2016, AIMS)	93
5.10	(Left) Some faces might not have landmarks ground truth. (Right) Some ground truth are not sufficiently accurate for comparison. (© 2016, AIMS)	93
5.11	Images on the left column are detected by Share-146 model (frontal model only), while the ones on the right column are detected by our proposed system. Share-146 miss the small faces and a false positive is detected on the background. (© 2016, AIMS)	94
5.12	(Left) Face region occupy a very small portion of the image. (Right) Faces occupy a large portion of the image (approximately 40%). (© 2016, AIMS)	95
5.13	Speed comparison on a face on small segment of the image (first scenario). (© 2016, AIMS)	95
5.14	Speed comparison on faces on large segment of the image (second scenario). (© 2016, AIMS)	96
5.15	Speed comparison after scaling down faces to 150x150 on MR-36 models (second scenario). (© 2016, AIMS)	97
6.1	Aligned face based on the proportion of eye centres and mouth centre. The face is then scaled to 360x320. (© 2015, IEEE)	101
6.2	Two chosen glasses shapes: (Top) Oval (Bottom) Rectangle. (© 2015, IEEE)	102
6.3	Our own created 39 glasses landmarks ground truth. (© 2015, IEEE)	103
6.4	The process of appointing 16 landmarks on a rim. (1) 4 landmarks on the left, right, top, and bottom position. (2) A landmark is added in the middle of each pair of previous landmarks. (3) Repeat the last procedure once more to pinpoint the final landmarks. (© 2015, IEEE)	103
6.5	Our proposed glasses models. The first model is an oval-shaped glasses while the other one is a rectangle-shaped glasses. (© 2015, IEEE)	104
6.6	Masks derived by (Top) linear interpolation and (Bottom) PieceWise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation. (© 2015, IEEE)	105
6.7	(Left) First layer of mask covering all base parts of glasses. (Middle) Additional layer of mask to cover nose pad, bridge, and lower rim. (Right) Combination of both layers of mask. (© 2015, IEEE)	105

6.8	2 state-of-the-art image reconstrution methods (NLCTV inpainting & SF-DAE Deep Learning model) structured in a "cascade" manner to filter the presence of glasses consecutively.	106
6.9	Our proposed glasses detection + removal system. (© 2015, IEEE)	107
6.10	This is the iterative process of NLCTV inpainting on glasses. The image on top left is the original face image with glasses. The next image shows the mask generated from the glasses landmarks which is then gradually reconstructed along with the glasses.	108
6.11	(Left) Aligned face images wearing glasses (Middle) Face images after NLCTV inpainting (Right) Face images after reconstruction via NLCTV inpainting + SFDAE Deep Learning model. All images in this example are contrast normalized through histogram equalization method and scaled down to 66x66. (© 2015, IEEE)	109
6.12	Since this is a rimless glasses, the edge features are too <i>faint</i> to consider it as a glasses. (© 2015, IEEE)	111
6.13	(Left) Original image without glasses (Middle) First synthetic data with thin silver glasses (Right) Second synthetic data with thick dark glasses. (© 2015, IEEE)	112
6.14	Illustration of experiment setup for our cross-identity testing.	114
6.15	Facial recognition with classification approaches PCA, LDA, and SRC. This result proves that removing presence of glasses improves the facial recognition rate.	115
6.16	ROC curves on thin glasses.	116
6.17	ROC curves on thick glasses.	117
7.1	The framework of our semantic-based face images retrieval. (1) We have 117 subjects with ten images each. Six of them contain various illumination which are normalized through Multi-scale Self Quotient image (MSQ) approach. 130 facial landmarks are then extracted from all the face images through our proposed component-based AR model. Furthermore, glasses presence labels are also created by detecting it through our proposed glasses model. (2) Geometric features (e.g eye distance and shape Triangular Area Region (TAR) feature) are extracted based on the facial landmarks information. All these features are mapped semantically to define their "membership degree" to each semantic benchmark sample. (3) These membership degree and glasses labels are then used as features to perform face images retrieval.	122
7.2	Samples of faces from AR database. All illuminated faces are normalized via Multi-scale-Self-Quotient-image (MSQ) approach.	123

7.3	(Top) Eyes shape/size are not necessarily affected by facial expressions. The eyes might look narrow or wide open on any facial expression. (Bottom) Our proposed component-based face models extended from AR model. Landmark fitting for both eyes are not affected by the facial expression on the lower part of the face.	125
7.4	All face images are categorized based on the presence of glasses.	127
7.5	(Top) The distance between both inner eye corners. We selected three types of benchmarks: close, medium, and far with respect to the width of the eye. (Bottom) The size of the eyes calculated through the ratio between its height and width. We have three types of eye size: narrow, medium, and widely-opened for both left and right eyes.	128
7.6	Three chin shapes.	130
7.7	Four right and left eye shapes.	131
7.8	Five right and left eyebrow shapes.	132
7.9	Six mouth shapes.	133
7.10	Six nose shapes.	134

List of Tables

3.1	Relative error and detection rate from Independent-1050 and AR model. (© 2014, IEEE)	53
3.2	Width and height error rate from Independent-1050 and AR model. (© 2014, IEEE)	53
3.3	Relative error and detection rate from CompASM and AR model. (© 2014, IEEE)	55
3.4	Width and height error rate from CompASM and AR model. (© 2014, IEEE)	55
3.5	Relative error and detection rate from CompASM and AR model on each expression. (© 2014, IEEE)	55
3.6	Relative error and detection rate of the AR model on various colour spaces. (© 2014, IEEE)	57
4.1	The summary of the MR models. All of them are trained on four facial expressions (neutral, smile, angry and scream) from 112 subjects from AR database. (© 2014, IEEE. 2016, AIMS)	67
4.2	11 Facial Landmarks Relative Error from the SHARE-146 model and MR Models. (© 2014, IEEE. 2016, AIMS)	73
4.3	Detection Rate (%) from the SHARE-146 Model and MR Models. (© 2014, IEEE. 2016, AIMS)	73
5.1	True Positive and False Positive on AFLW Database from the SHARE-146 Model (all 13 poses and single pose) and MR Models. (© 2016, AIMS) . .	90
6.1	Information on chosen face images on various databases. (© 2015, IEEE) .	110
6.2	Glasses detection rate on various databases. (© 2015, IEEE)	111
6.3	Average Euclidean distance between the synthetic data and neutral frontal face images. (© 2015, IEEE)	112
6.4	Face verification rate at 0.1% False Acceptance Rate (FAR) before and after glasses removal.	118
7.1	Facial landmarking performance improvement with component-based AR model.	124
7.2	Result of learning semantic combination on the first 50 subjects based on the success rate. These are evaluated with $n = 1, 2, 3, 4, 5$ queries.	139
7.3	The retrieval success rate on the remaining 67 subjects.	139
7.4	Success rate improvement before and after glasses filter.	140

7.5	Result of learning semantic combination on the first 50 subjects based on the success rate. These are evaluated with $n = 1, 2, 3, 4, 5$ queries.	141
7.6	The retrieval success rate on the remaining 67 subjects.	141

Chapter 1

Introduction

Human face is considered as one of the most popular biometric information of a person (Zhao *et al.*, 2003). Unlike other biometrics such as fingerprint and iris, face images can be captured unnoticed without individual's cooperation. This makes it much easier to collect large amount of data on face images. This is also supported by the vast development of image-acquisition devices such as surveillance cameras (e.g Closed-Circuit Television (CCTV) camera) and portable devices (e.g mobile phone). Some possible field applications include but not limited to surveillance/law enforcement, entertainment, and information security.

Human face is an abstract and complex feature containing a vast amount of information for various purposes. From the face, we can learn the identity of the person (face recognition/verification), facial expressions/emotions, or even the intention based on the gaze. One of the practical applications is the face image retrieval where we retrieve images containing the face(s) of the query subject. Some of the recent developments were proposed by Conilione and Wang (2012), Li *et al.* (2015), Arandjelovic (2016), and Bhattarai *et al.* (2016). However, before we can extract all these information from the face, we have to get the answer to the questions: "Is there any face in this image? If yes, where are they? How many are there?". The answers for all these problems are addressed in the field of **Face Detection**. According to Zhang and Zhang (2010), one of the most popular face detection approaches is the Viola and Jones face detector (Viola and Jones, 2004) due to its efficient and robust performance. It is the result of combination of a novel image representation *integral image*, Haar features (Papageorgiou *et al.*, 1998), and Adaboost learning algorithm (Freund and Schapire, 1995) in a cascade structure framework.

Detecting locations of the faces in an image is not always sufficient in many applications. Usually, we need supplementary information such as the location of facial components (e.g eyes, nose, mouth). This scenario appears on face recognition. Although some techniques process the face region holistically (Liu and Wechsler, 2000; Li and Lu, 1999; Bartlett, 2001), other approaches require local features from the facial components (Okada *et al.*, 1998; Nefian and Hayes III, 1998; Lawrence *et al.*, 1997). Some other techniques even combine both holistic and local features as hybrid approaches (Huang *et al.*, 2003; Penev

and Atick, 1996; Lanitis *et al.*, 1995). Therefore, it is necessary to process the face images further to extract these local features. This problem is usually known as the **Facial Landmarking** or **Facial Landmarks Detection** (Çeliktutan *et al.*, 2013). It is defined as the process to automatically localize particular characteristic points/landmarks on faces, which is a necessary phase to many face processing applications. However, this task is proven to be challenging due to various factors such as face poses, facial expressions, illumination and occlusions. All the techniques for extracting facial landmarks from face images are usually divided into two categories: texture-based and model-based. Some of the recent examples are (Valstar *et al.*, 2010; Ding and Martinez, 2010; Akakin and Sankur, 2007) for texture-based and (Zhu and Ramanan, 2012a; Belhumeur *et al.*, 2013; Milborrow and Nicolls, 2008) for model-based. According to Çeliktutan *et al.* (2013), the model-based techniques usually perform better than the texture-based techniques. Additionally, there are also some techniques specifically designed for 3D faces such as (Nair and Cavallaro, 2009; Dibeklioglu *et al.*, 2008; Akagunduz and Ulusoy, 2007). However, 3D faces are not the research focus of this thesis.

Among all the facial landmarking techniques, the one that stands out the most is the technique proposed by Zhu and Ramanan (2012a) due to its capability to perform face detection, facial landmarking and pose estimation simultaneously with reliable performance. Therefore, we utilize the concept of pictorial-tree-structured face models proposed in (Zhu and Ramanan, 2012a) as the framework foundation for all the proposed approaches in this thesis.

1.1 Research Gaps and Aims

After conducting in-depth investigations on the literature review, we discovered a few research gaps to be addressed as follows:

1. Many facial landmarks detection approaches perform quite well in localizing the landmarks on facial components. However, the amount of detected landmarks are usually restricted to the intended applications (Çeliktutan *et al.*, 2013). For instance, face recognition might only need approximately 20-30 landmarks just to enclose the facial components in a bounding box. The more complex tasks such as facial expressions understanding and facial animation might need up to 60-80 landmarks for high accuracy. Normally, facial landmarks detection approaches do not extract much higher amount of landmarks due to the extra computational cost. However, the application on semantic-based face images retrieval will need much more landmarks

to derive complex features such as the shape of the eyes (e.g (Conilione and Wang, 2012)). As far as we know, there is no facial landmarks detector designed for this particular problem. Most semantic-based face images retrieval frameworks rely on manually-assigned facial landmarks or facial landmarks detected with low amount of landmarks to derive only very basic geometric features. This is why we want to develop a facial landmarks detector with significantly large number of landmarks and better accuracy.

2. Face images might come in various resolutions in the image. While high resolution faces usually do not pose any problem, the low resolution faces might cause a problem with its limited information. Some facial landmarks detector approaches can still be applied on low resolution faces but with the same amount of landmarks. We believe it is not a good idea to crumple high amount of landmarks on the small faces (e.g 30x30) as it might disrupt the structure of the landmarks. There are also facial landmarks approaches which are not able to perform on small faces due to the difference on features between large faces and small faces. This is why we want to develop face models which can extract facial landmarks on various face resolutions accordingly with proper structure and amount of landmarks.
3. The Viola Jones face detector (Viola and Jones, 2004) is able to detect faces efficiently with high face detection rate. However, our observation discovers that their approach is still susceptible to large amounts of false detection in uncontrolled environment. Even though the face models proposed by Zhu and Ramanan (2012a) have been proven to have better face detection performance than the Viola Jones face detector, the computational cost is far from real-time due to its simultaneous facial landmarks detection process. This is why we want to develop a new technique which can perform with high face detection rate and low false detection in relatively short time in uncontrolled environment.
4. In the context of semantic-based face images retrieval, facial components such as eyebrows, eyes, nose and mouth are not the only semantic features we can extract from a face. The presence of the glasses can also be considered as the main component of the face due to its high usage among people for either visual problems or fashion as mentioned by Gao *et al.* (2008). All the current facial landmarks detector we are aware of never consider to detect the glasses. This is why we want to develop a novel tree-structure model for glasses landmarks detection. Furthermore, since the presence of glasses affect the facial recognition performance negatively (Righi *et al.*, 2012), we also propose a framework to remove the presence of glasses to improve both facial recognition and verification performance.

1.2 Thesis Structure and Contributions

The list below briefly describes the content of each chapter in this thesis along with its contributions (Figure 1.1).

- Chapter 2: We introduce some **preliminary knowledge** related to our proposed approaches in this thesis. We begin with introducing the concept of pictorial-tree-structure face models (Zhu and Ramanan, 2012a) along with its gradient-based HOG features (Dalal and Triggs, 2005). It is then followed by the explanation of the widely-used Viola Jones face detector (Viola and Jones, 2004). Furthermore, two state-of-the-art image reconstruction approaches, the NLCTV inpainting (Duan *et al.*, 2015) and SFDAE Deep Learning model (Pathirage *et al.*, 2015) are explained. Moreover, some commonly used face recognition techniques: the PCA (Turk and Pentland, 1991), LDA (Belhumeur *et al.*, 1997), and SRC (Wright *et al.*, 2009) are described in detail. Finally, we present brief descriptions and images samples for all facial/non-facial images database used in this thesis.
- Chapter 3: We propose the novel high-density frontal face models called the **AR model**. We design a new face structure emphasized heavily on eyebrows, eyes, nose, and mouth. This model is able to detect 130 facial landmarks, almost twice as many as the previous state-of-the-art pictorial-tree-structure face models Independent-1050 (Zhu and Ramanan, 2012a). The advantage of the large amount of facial landmarks allows us to describe better semantic features of facial components. Experimental results reveal the significant improvement on both accuracy and detection rate on fiducial points achieved by the AR model against some other state-of-the-art facial landmarking techniques. Additionally, it also shows higher accuracy on defining basic semantic features on facial components. Lastly, we conduct a full investigation on the impact of various colour spaces on facial landmarks detection with our proposed AR model.
- Chapter 4: We present the **Multi Resolutions (MR)** face models for performing facial landmarks detection on low resolution faces as small as 30x30 on which the predecessor state-of-the-art face models Share-146 (Zhu and Ramanan, 2012a) would fail. For the purpose of assisting the face models training, we design an automatic **adaptive landmark scheme** for facial landmarks selection on various resolution levels of the face. This allows us to train face models on any resolution with sufficient amount of landmarks. Furthermore, in order to utilize the MR models more effectively, we employ the Viola Jones face detector (Viola and Jones, 2004) prior to

facial landmarking phase. This setup let us decide which face model scale to apply automatically. Experiments are carried out on faces with various scale levels with a few state-of-the-art techniques. The results emphasize the robust performance of our proposed MR models on high resolution and stability on low resolution especially in the presence of beard and hair.

- Chapter 5: We propose a novel face detector method with the **Tree-structured Filter Model (TFM)**. This model filters all the face regions detected by the Viola Jones face detector (Viola and Jones, 2004) to remove most false detections. In order to avoid high overhead from the additional processing, TFM is designed to be light-weight by training it on the low resolution faces just sufficient to depict the intuitive description of human faces. The experiments are conducted on an uncontrolled face database which reveals the advantage of TFM in terms of computation speed and detection rate compared to the Viola Jones face detector and another state-the-art face detection model. We also design a complete integrated framework of facial landmarking system by combining it with the previously proposed MR models. More experiments reveal that the integrated system performs better on uncontrolled environment and not significantly affected by the size of the image.
- Chapter 6: We investigate the feasibility of utilizing the concept of pictorial tree structure for proposing novel **glasses models** for automatic detection of glasses presence along with its corresponding landmarks on face images. We address this problem by training the tree-structure model on 100 glasses images with the corresponding 39 manual landmarks. The landmarks are created systematically to ensure the consistency and accuracy of the landmarks. We integrated this model with two state-of-the-art image reconstruction approaches NLCTV inpainting (Duan *et al.*, 2015) and SFDAE Deep Learning model (Pathirage *et al.*, 2015) as a novel double-layers glasses filter framework to automatically remove the glasses in order to improve the facial recognition and verification rate. Our experiments reveal the robustness of our proposed glasses models on detection rate on various face databases. Furthermore, it also confirms the significant improvement caused by glasses removal on face recognition and verification.
- Chapter 7: We develop an **automatic semantic-based face images retrieval** integrated with the proposed **AR model** and **glasses models** to derive the semantic features from face images. We make an adjustment on AR model to be a component-based face model in order to further improve its accuracy and less influenced by slight facial expressions. The new proposed AR model consists of three independent partial face structures: (1) left eyebrows and eyes (2) right eyebrows and eyes (3)

lower region on the face including nose, mouth, and chin. Thus, we can automatically extract all the facial landmarks from all the face images in the gallery dataset and derive the semantic features (e.g shape and size) from them. Furthermore, we provide a wide range of semantic benchmarks chosen manually to define some categories for each semantic feature (e.g narrow, medium, and widely-opened eye). These benchmarks are used to apply *semantic mapping* for all extracted semantic features on each face. This process will assign the "membership degree" as features for each face. The advantage of this approach is that it is efficient and only involves small-scale data assigned manually by hand. The experiment results reveal the practicality of semantic features for face images retrieval with high success rate of finding the correct identity of the query subject.

- Chapter 8: The **conclusion** of the whole thesis and some **potential future directions** are addressed.

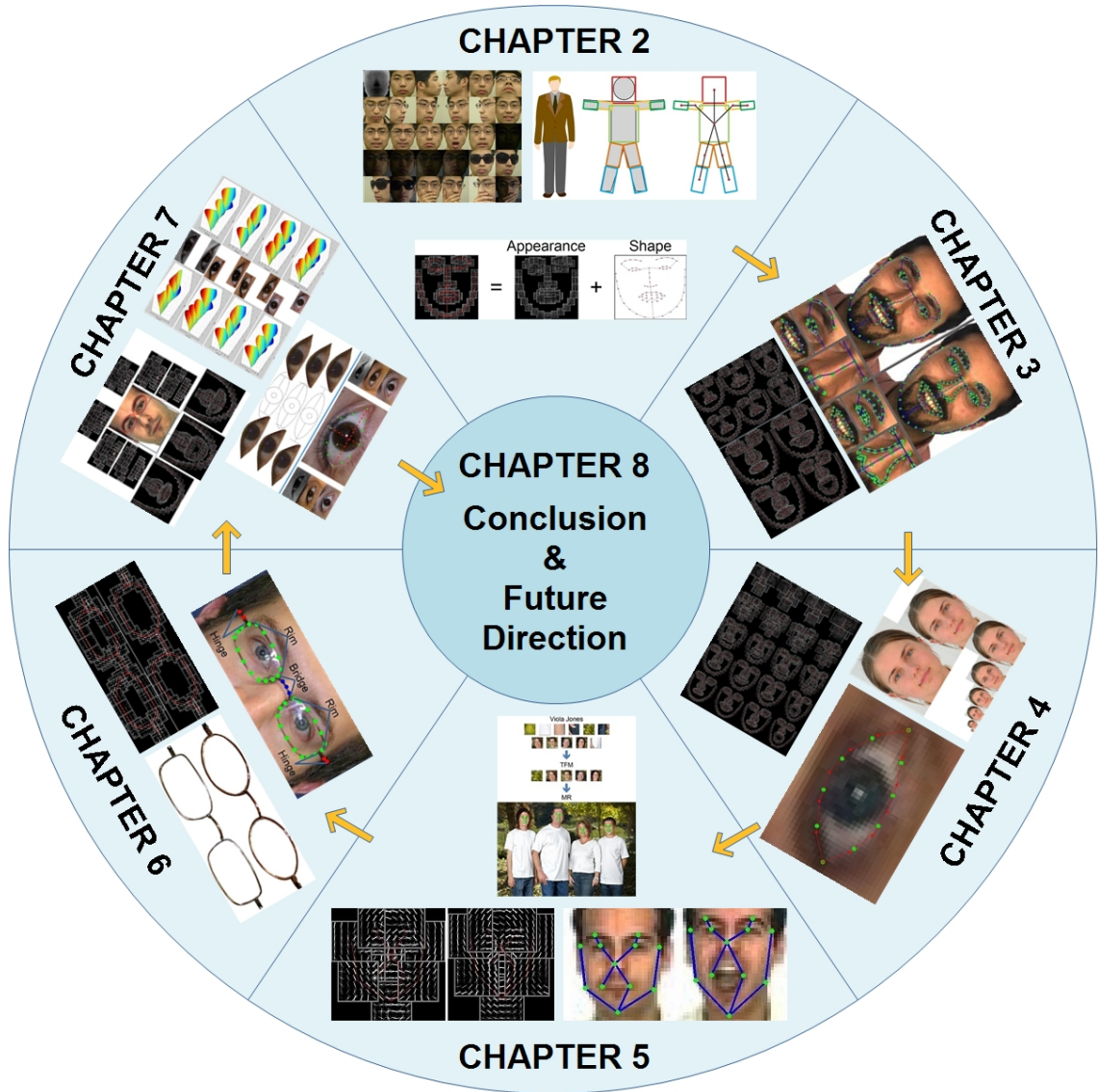


Figure 1.1: Visual cues of the remaining chapters. Starting from the top moving clockwise, we briefly discuss about the related approaches/techniques and relevant databases used in our experiments in Chapter 2. Our research is mainly focused on pictorial-tree-structured models to detect landmarks on facial region with various implementations. In Chapter 3, we proposed robust frontal face models with higher accuracy and condensed landmarks in order to extract semantic features from the human faces. We further expand the capability of our face models in Chapter 4 to cover various resolutions as low as 30x30. In order to have an efficient facial landmarks detection framework, we propose light-weight face models for fast initial face detection in Chapter 5. In Chapter 6, we propose novel glasses landmarking models as facial glasses can be considered as a part of the face and its presence is highly common. Lastly, we apply our proposed face and glasses models to extract facial semantic features and conduct facial images retrieval in Chapter 7. The conclusion and future directions are summarized in Chapter 8

Chapter 2

Background

In this chapter, background knowledge and related techniques used in this thesis are presented. All the databases involved on all our experiments are also described. The contents are summarized at the end of this chapter.

2.1 Pictorial-Tree-Structured Face Models

The foundation of the works presented in this thesis is based on the state-of-the-art approach proposed by Zhu and Ramanan (2012a). Their approach was designed to accomplish multiple tasks in one integrated pictorial-tree-structured-based framework. These tasks are the face detection, face pose estimation, and face landmarks detection. This means that there are no prior information such as location of the faces or the amount of the faces in the image required. This framework can be applied on any images without any restriction (uncontrolled environment).

The robustness and flexibility of the approach come from their proposed face models. Each face model is derived from a mixture of facial landmarks connected as a pictorial tree structure (Felzenszwalb and Huttenlocher, 2005) which is suitable for preserving the global elastic formation of the faces. The feature extracted from each facial landmark is the Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005). These features describe the orientation of edges on a local region in an image by calculating the distribution of intensity gradients.

2.1.1 Pictorial Structure

The original concept and framework of pictorial structure models were introduced by (Fischler and Elschlager, 1973). This was further developed by (Felzenszwalb and Huttenlocher, 2005) for the purpose of recognizing any general object by conducting experiments on human faces and bodies recognition. Intuitively, the idea of pictorial structure is that

an object is represented by a collection of parts of interest (features) connected in a particular framework/structure which represents the relation between them. To get a better intuitive idea, refer to an example in Figure 2.1.

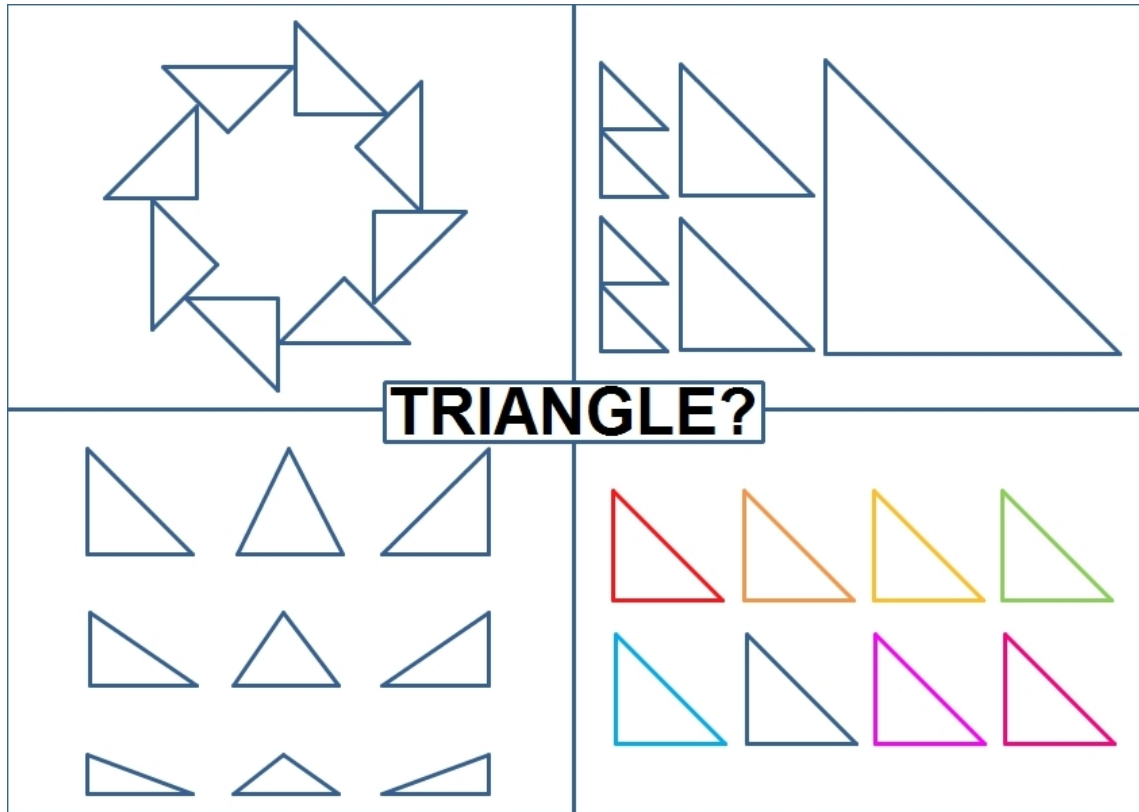


Figure 2.1: These are 2D triangle shapes in various orientations, sizes, scales/ratios, and colours. Even with these large varieties of deformation, people still recognize the triangle shape.

This is a collection of 2D triangle shapes in various deformations. Despite the enormous number of varieties of triangles, we still can easily recognize the shape. The reason is not because we memorize all possibilities of triangles, but because we have learned the fundamental characteristics of a triangle. We observed that each shape in Figure 2.1 is a closed shape with only 3 corners connected via 3 intersecting lines. In this scenario, the corners of the triangles are the "parts" and the "relation" between them are represented by the three straight lines composing the shape. Another more sophisticated example is the pictorial structure on human (standing straight) as can be seen in Figure 2.2.

A human body can be divided into multiple body "parts" (head, torso, upper arm, lower arm, upper leg, lower leg). The presence of each part can be described by particular features (e.g colour, shape, edges). A "relation" can be defined on these body parts.

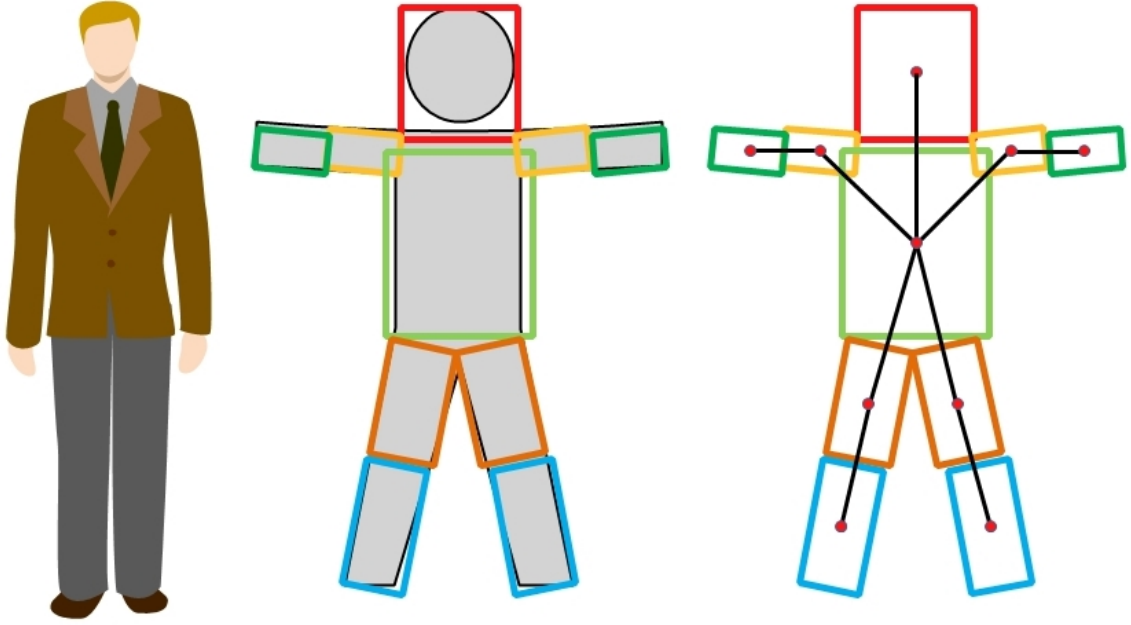


Figure 2.2: A pictorial structure of a standing human. The whole body can be represented by ten parts: head, torso, 2 upper arms, 2 lower arms, 2 upper legs, 2 lower legs.

For instance, we can intuitively describe that the human head is adjacent to torso (close distance) and located on the upper part (relative position). These configurations of parts definition and the corresponding relationship can describe a presence of human in general.

The formal definition of a pictorial structure can be expressed as a undirected graph $G = (V, E)$. $V = v_1, v_2, \dots, v_n$ represents the "parts" of an object. Each part has a corresponding "configuration" variable $L = l_1, l_2, \dots, l_n$. Normally, L just represents a location (x, y) of a part v_i in the image. However, additional information can be added such as the orientation. Each direct connection between two parts is represented by an edge $(v_i, v_j) \in E$.

Once the object model has been learned, the matching can be regarded as a minimization problem of the cost function between the model and a region in the image. There are two main parts in the cost function. First, it evaluates the degree of mismatch between part v_i at location l_i in the image with a function $m_i(l_i)$. Second, it measures the degree of deformation between parts v_i and v_j with a function $d_{ij}(l_i, l_j)$. The matching formula can be written as:

$$L = \arg \min_L = \left(\sum_{i=1}^n m_i(l_i) + \sum_{(v_i, v_j) \in E} d_{ij}(l_i, l_j) \right) \quad (2.1)$$

2.1.2 Histogram of Oriented Gradients (HOG)

The face models proposed by (Zhu and Ramanan, 2012a) demand a feature which can describe the local region of each facial landmark well. One of the suitable features is the one based on the concept of orientation histograms developed in the early age (McConnell, 1986; Freeman and Roth, 1995; Freeman *et al.*, 1996). This concept reached a significant performance improvement by involving local spatial histogram and normalization in image descriptor SIFT (Scale Invariant Feature Transform) (Lowe, 2004). Zhu and Ramanan adopted the HOG (Histograms of Oriented Gradients) features by (Dalal and Triggs, 2005) as it depicts the appearance and geometric information well. For instance, each facial landmark on the chin region represents a local silhouette along the jawline. This is achieved by exploiting the information on the magnitudes and orientations/directions of the intensity gradients distribution.

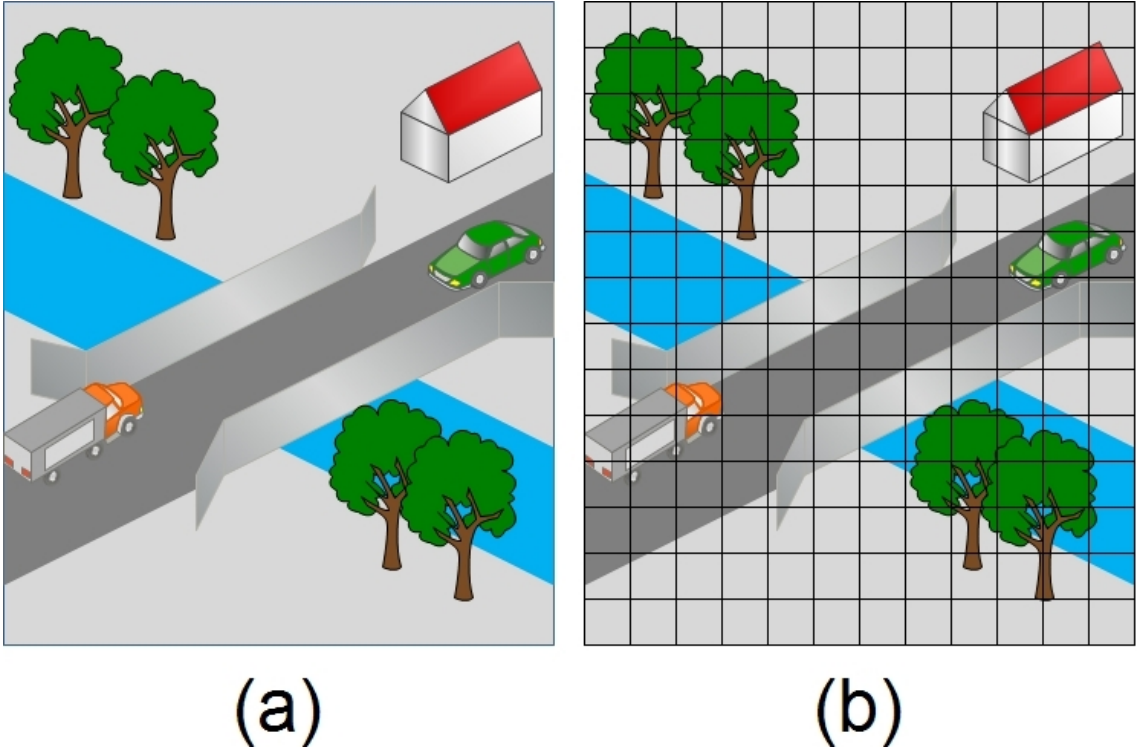


Figure 2.3: (a) Original image (b) Rectangular partitions of the image called "cells".

In order to get a better intuitive concept of HOG features extraction, a simple simulation is briefly explained in this section. Assuming we have a colour image $I \in \mathbb{Z}^{180 \times 210 \times 3}$. The image is then divided into rectangular partitions called "cells" as can be seen in Figure 2.3. The intensity gradients is extracted for each cell. The way to achieve this is by measuring the pixel values difference on the horizontal (d_x) and vertical (d_y) directions. Refer to Figure 2.4 for the calculation done by applying a centered mask for each direction. As a result, it is feasible to measure the magnitude $mag = \sqrt{(d_x)^2 + (d_y)^2}$ and orientation $\theta = \arctan(\frac{d_y}{d_x})$. If the image contains multiple colour channels (e.g RGB (Red, Green, Blue)), then the calculation is done separately on each channel for each pixel. The one with the largest magnitude is chosen.

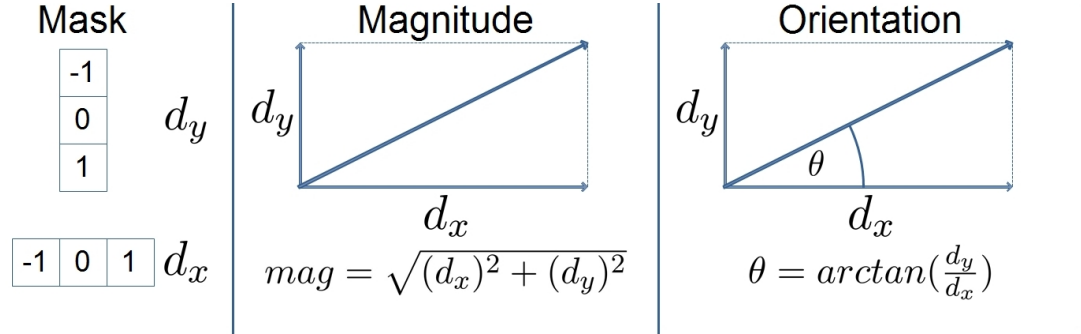


Figure 2.4: Pixel intensity differences can be calculated by applying masks on both horizontal (d_x) and vertical (d_y) directions. Both values can be used to compute the magnitude/strength and the orientation/angle (θ) of the gradients.

All the magnitudes and orientations are accumulated to create a small-scale histogram on various angles. However, this might lead to a huge dimension of data. For instance, even after rounding to the nearest integer, there are still 180 angles to be considered ($1^\circ, 2^\circ, 3^\circ, \dots, 180^\circ$). In order to prevent the abundance of data dimension, the orientation has to be quantized into evenly space based on the number of **spatial/orientation bins**. (Dalal and Triggs, 2005) used 9 orientation bins which divides the angles with the increment of $\frac{180}{9} = 20^\circ$ degrees ($20^\circ, 40^\circ, 60^\circ, \dots, 180^\circ$). If the extracted orientation does not fall exactly into one of these angles (e.g 45°), then the magnitude is linearly interpolated between the neighboring bin centers (e.g 75% into 40° and 25% into 60°).

Finally, the value of multiple cells can be combined into a block which represents a larger comprehensive representation of the HOG features (Figure 2.5). Describing the whole image can be done by concatenating all the blocks (data dimension = amount of blocks * amount of cells per block * spatial/orientation bins).

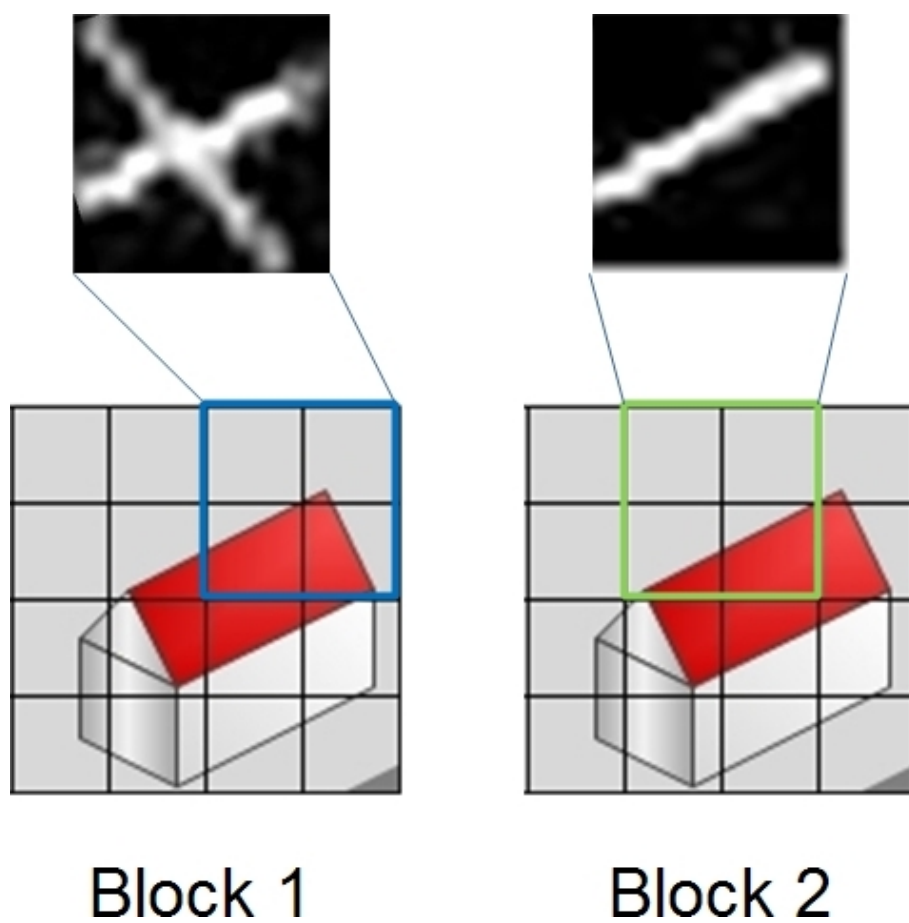


Figure 2.5: Examples of blocks with 2x2 cells. Block 1 covers the edge of the roof with 2 distinct orientations which is shown as 2 lines in the corresponding HOG feature. Block 2 only contains 1 orientation.

2.1.3 Zhu and Ramanan Face Models

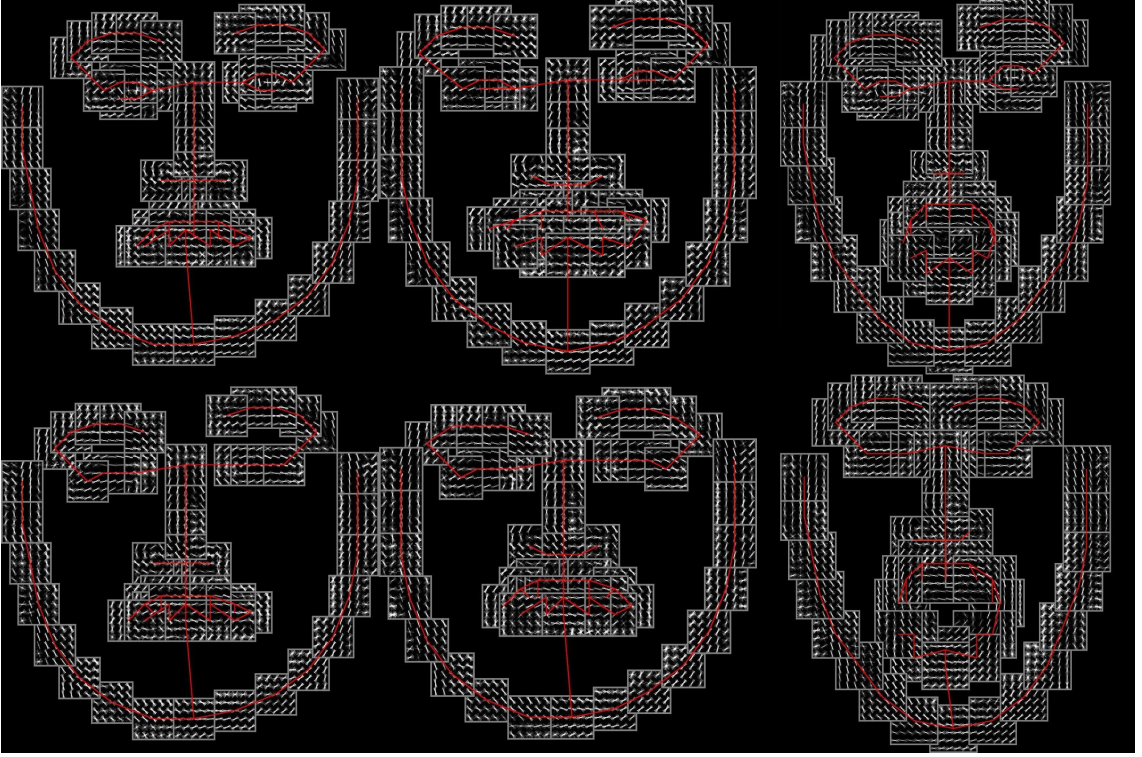


Figure 2.6: 6 mixtures of tree structure for 6 variations of facial expressions (neutral, smile, surprise, squint, disgust, scream) from CMU multiPIE database (Gross *et al.*, 2010). (© 2014, IEEE)

Zhu and Ramanan (2012a); Yang and Ramanan (2011) extend and apply the idea of pictorial structure further to detect the presence of human faces and provide the facial landmarks based on the optimal configuration L . Their face models consist of m mixtures of tree structure to represent various poses and facial expressions. For instance, 6 mixtures of tree structure are needed to express 6 facial expressions (Figure 2.6). The reason of employing tree architecture is because it can be optimized through dynamic programming (Felzenszwalb and Huttenlocher, 2005). Let $T_m = (V_m, E_m)$ indicate a pictorial tree structure of mixture m with a collection of "parts" V and the corresponding "relations" E (similar concept described in chapter 2.1.1). Each part is accompanied with a configuration $l_i = (x_i, y_i)$ which specifies pixel location of part i . All configurations are defined as $L = \{l_1, l_2, \dots, l_i : i = |V|\}$. They compute the score of the configuration matching in an image I as follows:

$$Score(I, L, m) = Appearance_m(I, L) + Shape_m(L) + \alpha^m \quad (2.2)$$

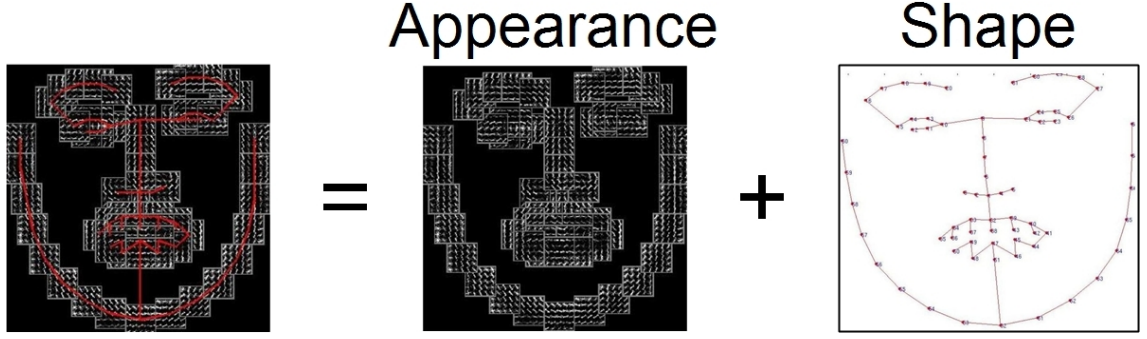


Figure 2.7: Each face model involves 2 main components. The first component is the appearance which describes the HOG features (Dalal and Triggs, 2005) on each part of the tree structure. The second component is the comprehensive arrangement of the relations among all the parts which defines the overall tree shape of the face.

Basically, the computation involves two crucial components: the **appearance** evidence from the learned features and **shape** arrangement of the structure of L . α^m is the scalar bias of mixture m . The visual definition can be seen in Figure 2.7. Both components are defined as:

$$Appearance_m(I, L) = \sum_{i \in V_m} w_i^m \cdot \phi(I, l_i) \quad (2.3)$$

$$Shape_m(L) = \sum_{i, j \in E_m} a_{ij}^m dx^2 + b_{ij}^m dx + c_{ij}^m dy^2 + d_{ij}^m dy \quad (2.4)$$

Eq. (2.3) calculates the whole amount of appearance indications of all the learned templates w_i^m for configuration l_i on mixture m compared to the features $\phi(I, l_i)$ extracted on the location l_i of image I . A strong indication implies that the region more likely contains a human face. However, a decent feature matching is still inadequate to form a conclusion. The configuration L has to match the spatial arrangement of a face well as computed in Eq. (2.4). $dx = x_i - x_j$ and $dy = y_i - y_j$ define the horizontal and vertical displacements of the relative position between connected parts i and j . Parameters (a, b, c, d) control the intensity of each term. As there are multiple mixtures m to be matched on an image, the inference process will be determined by choosing the one which generates the highest configuration score:

$$Score^*(I) = \max_m \left[\max_L Score(I, L, m) \right] \quad (2.5)$$

All the face models were trained in a supervised manner with both positive and negative images. Positive images contain human faces along with the associated facial landmarks ground truth and mixture labels. On the other hand, negative images consists of non-facial images. These data were used to learn both shape and appearance parameters. An approach proposed by (Chow and Liu, 1968) was employed to discover the maximum likelihood tree structure to estimate E_m . With the learning approach by (Yang and Ramanan, 2011), let $z_n = \{L_n, m_n\}$ be the configuration and mixture of the positive images, if all parameters (w, a, b, c, d, α) are grouped together as a vector β , the score function can be defined as:

$$S(I, z) = \beta \cdot \Phi(I, z)$$

with nonzero elements in $\Phi(I, z)$ corresponding to mixture m . The model can be learned from:

$$\begin{aligned} & \arg \min_{\beta, \xi_n \geq 0} \frac{1}{2} \beta \cdot \beta + C \sum_n \xi_n \\ \text{satisfying } & \begin{cases} \beta \cdot \Phi(I_n, z_n) \geq 1 - \xi_n & , \quad \forall n \in \text{positive samples} \\ \beta \cdot \Phi(I_n, z) \leq -1 + \xi_n & , \quad \forall n \in \text{negative samples}, \forall z \end{cases} \\ & \text{and} \\ & \beta_k \leq 0, \quad \forall k \in K \end{aligned}$$

where the positive samples will generate high score (≥ 1) and negative samples will generate low score (≤ -1) with violation penalty variable ξ_n . K represents the parameters a and c in vector β .

2.1.4 Source Code and Pre-trained Models

As part of their research, Zhu and Ramanan (2012a) provide the open source code for both model training and testing in (Zhu and Ramanan, 2012b). Furthermore, they also provide a few pre-trained face models. The first face model is the *Independent-1050* which is the most extensive and comprehensive model because it consists of 18 mixtures (Figure 2.8). 13 mixtures cover various face poses from -90° to 90° with 15° increment including neutral frontal face. 5 more mixtures were added to express 5 distinct frontal facial expressions (smile, surprise, squint, disgust, and scream). Not all the mixtures

contain equal amount of landmarks. Both frontal and near-frontal faces (-45° to 45°) include 68 landmarks while the profile faces (-90° to -60° and 60° to 90°) only involve 39 landmarks because of the invisible parts of the face. In total, *Independent-1050* contains $(68) \cdot (12) + (39) \cdot (6) = \mathbf{1050}$ landmarks on which each has its own HOG descriptor. Each landmark is a collection of 5×5 HOG cells with spatial bin of 4.

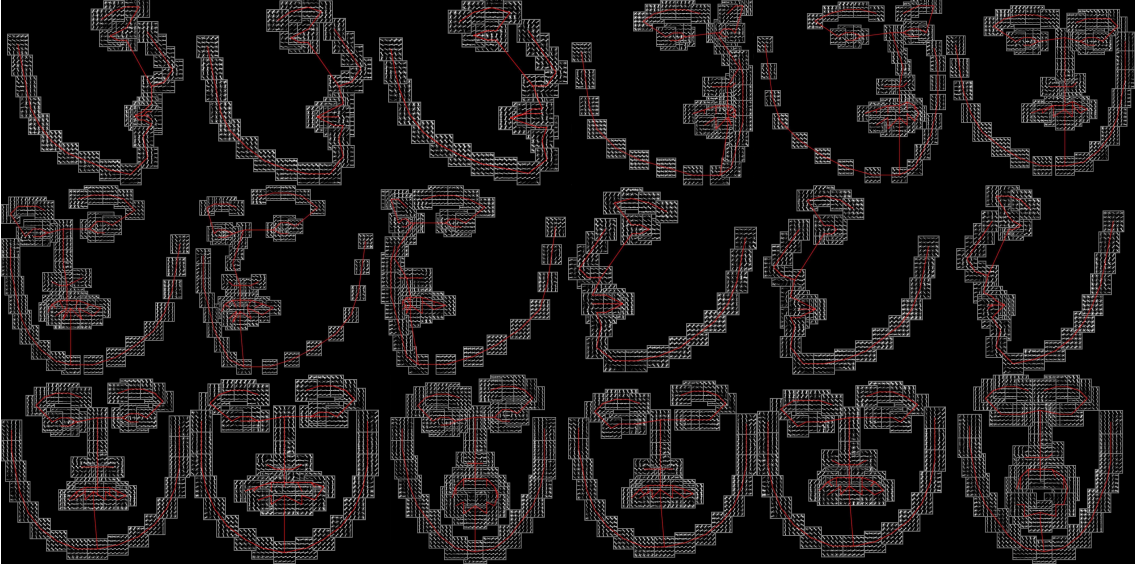


Figure 2.8: *Independent-1050* face models proposed by (Zhu and Ramanan, 2012a) with 18 mixtures. It covers 13 facial poses including frontal poses and 5 facial expressions.

Inspired by (Torralba *et al.*, 2007), Zhu and Ramanan’s face models allow sharing descriptors between similar parts. For instance, the eye corners of frontal face (0°) are quite similar to the one on 15° and 30° faces. This leads to a faster and more efficient model since there are fewer features to match with a consequence of a slight reduction in landmarking accuracy. Zhu and Ramanan provide two other face models which share the HOG descriptor. The first one is the extensive shared model called the *Share-99* where each particular facial landmark only has one HOG descriptor among all the mixtures. In total, only 99 HOG descriptors are adequate. The second model (the *Share-146*) is more flexible where it shares the parts only if the model has a similar topology. The varieties of the mixtures for this model are divided into 3 categories: frontal/near-frontal faces (-45° to 45°) and 2 profile faces (right/left) (-90° to -60° and 60° to 90°). As mentioned previously, frontal/near-frontal faces contain 68 landmarks while the profile faces contain 39 landmarks. Overall, the *Share-146* consists of $(68 + 39 + 39 = 146)$ unique descriptors. These two part-sharing models only cover 13 mixtures of facial poses without facial expression other than neutral.

All these face models were trained with face images collected from CMU multiPIE database

(Gross *et al.*, 2010) as positive samples and non-facial images from INRIA person database (Dalal and Triggs, 2005). 50 face images were used to train one individual mixture of face. 650 face images cover 13 facial poses (-90° to 90°) and other 250 images cover 5 various facial expressions (*Independent-1050* only). On the other hand, all 1218 non-facial negative images were chosen which contain various objects such as but not limited to building, sky, road, and mountain.

Another significant advantage of Zhu and Ramanan’s proposed technique is the low requirement of training data availability. An extensive analysis conducted by Zhu *et al.* (2012) shows that the pictorial-tree-structure face models can be trained optimally even with low number of positive training data. They claimed that the minimum of 50 face images is sufficient to achieve high performance. Additional training data are unnecessary, but it can improve the performance slightly.

2.2 The Viola and Jones Face Detector

Viola and Jones (2004) proposed a simple yet robust and efficient face detector. Their approach is one of the most used face detector approaches and well-known in the field of face image analysis and understanding (Zhang and Zhang, 2010). Three main contributions for this detector are as follows. First, a new image representation called the "integral image" was proposed in order to compute the Haar-like features (Papageorgiou *et al.*, 1998) with time complexity of $O(1)$ (constant time). Second, Adaboost learning (Freund and Schapire, 1995) was applied to choose only few crucial features from an enormous amount of Haar features to build an efficient face classifier. Third, the architecture of the learned classifier was designed in a "cascade" manner which is capable of eliminating non-faces regions quickly and devoting more computation time on the promising face regions.

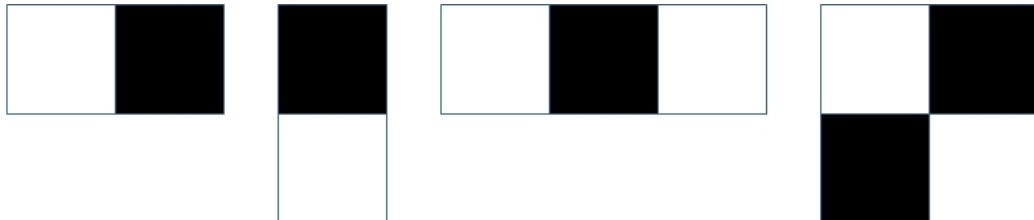


Figure 2.9: Some examples of Haar features used in Viola & Jones face detector.

The idea behind the Haar feature (Figure 2.9) is simple. For each Haar feature, calculate the difference of the sum of image intensities between white region and black region. This scalar value can reveal a faint indication of particular facial components in the face. For

example, region around nose bridge will have a significant image intensity difference from the sides while plain forehead region will produce low difference. All these features were computed in various image positions and scales/sizes. This leads to a prohibitively very high number of features which negatively impact the computation efficiency. This is the reason that motivates them to propose a new image representation "integral image" *int* derived from an image *im*:

$$int(x, y) = \sum_{i=1}^x \sum_{j=1}^y im(i, j)$$

This allows for a rapid computation of any Haar feature regardless of its position and scale (Figure 2.10). Calculation for the sum of image intensities can be done in a constant time.

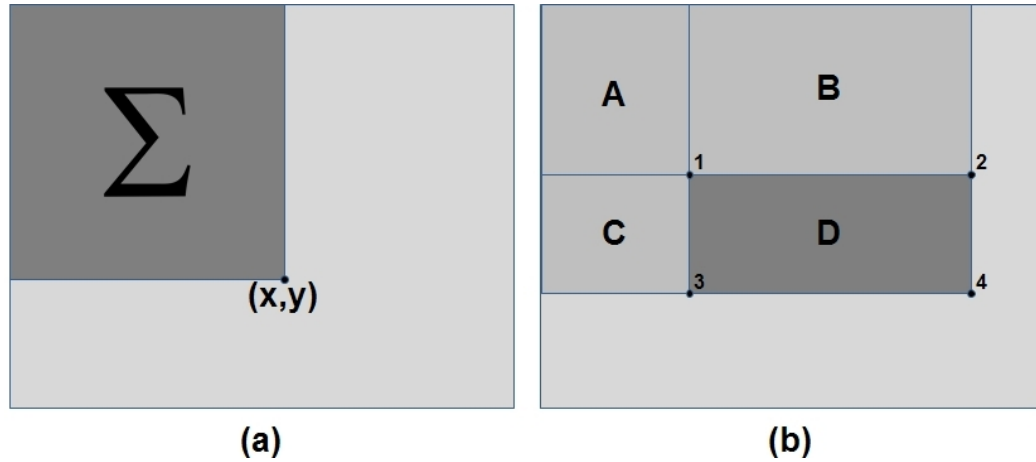


Figure 2.10: **(a)** "Integral image": the value of any position is the sum (Σ) of all pixel image intensities on the left **or** top of it. **(b)** Sum of image intensity on any region can be done by basic arithmetical operations on all four corner values. In this example, region *D* can be computed as $sum = D - C - B + A$.

Despite the speed enhancement via integral image, there is still a huge number of features to be processed for training a classifier. Viola and Jones employed the Adaboost learning algorithm (Freund and Schapire, 1995) to train a robust classifier by combining multiple weak classifiers. Each Haar feature is considered a weak classifier and only limited amount will be chosen to derive the strong classifier. Because of the features elimination process, intuitively this approach can also be considered a greedy feature selection technique. Some of the chosen relevant features are shown in Figure 2.11.

The combination of integral image and Adaboost produces a robust and efficient classifier for face detection. Despite the quick performance, there is still a concern about the high

number of input (image sub-window) to be processed. Apparently, most of the input are the non-faces regions (background or incomplete faces). Considering this fact, Viola and Jones restructured the architecture of the learned classifier in a "cascade" manner. By breaking the classifier into multiple stages, non-faces region can be eliminated quickly in the early phase thus spending more computation on the promising face regions (Figure 2.12).



Figure 2.11: (a) Original face image. (b) Haar feature indicates the presence of eyes by emphasizing the dark region on eyes and bright region on upper cheeks. (c) Haar feature indicates the presence of eyes by emphasizing the bright part on nose bridge covered by dark region on both sides.

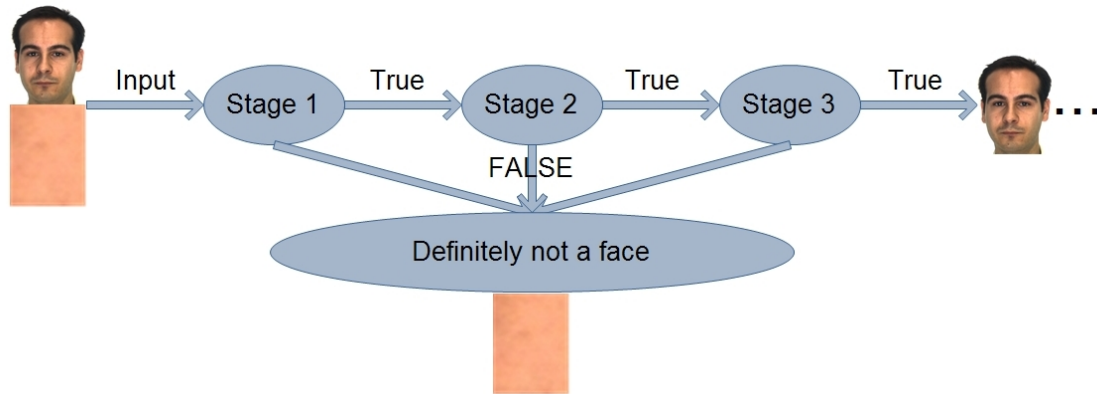


Figure 2.12: The classifier has multiple stages. If the input sub-window fails in any stage, it will be considered a non-face region. This allows for a quick elimination procedure and assigning more computation on the face regions.

2.3 Image Reconstruction Approaches

2.3.1 The NLCTV (Non-Local Colour Total Variation) Inpainting

The NLCTV (Non-Local Colour Total Variation) is one of the state-of-the-art inpainting technique by (Duan *et al.*, 2015). Inpainting is an image restoration technique which reconstructs the missing/corrupted region based on the other image region which remains intact, especially the surrounding/neighbour pixels (Bertalmio *et al.*, 2000). The assumption is that the location and the shape/size of the stained region is known beforehand. The total variation (TV) model (Shen and Chan, 2002) is favored by many researchers due to its simplicity and capability to recover sharp edges. However, the definition of the objective function in Bounded Variation (BV) space and its feature domain restriction on local information only lead to an unwanted staircase effect (i.e. patch/blocky appearances) and thus not suitable for recovering complex texture regions. This problem has to be solved by associating non-local information for better reconstruction. (Buades *et al.*, 2005) employed this solution to handle sophisticated texture. Their approach is extended further by (Gilboa and Osher, 2008) to define the non-local gradient, divergence, and other non-local operators by applying concepts of graph gradients and divergence. (Duan *et al.*, 2015) considered the application on the colour images by proposing coupling of multiple colour channels. This preserves the structure and texture of the face while recovering the missing region. The variational model of NLCTV is described as:

$$\min_u \{ E(u) = \sqrt{\sum_{i=1}^n \left(\int_{\Omega} |\nabla_{NL} u_i|(x) dx \right)^2} + \frac{1}{2} \sum_{i=1}^n \int_{\Omega} \lambda_D (u_i - f_i)^2 dx \} \quad (2.6)$$

where $f = (f_1, f_2, \dots, f_n)$ and $u = (u_1, u_2, \dots, u_n)$ are the original and recovered image respectively. $D \subset \Omega$ denotes the missing/broken region to be inpainted which is represented by a mask function $\lambda_D(x)$:

$$\lambda_D(x) = \begin{cases} 0 & x \in D \\ 1 & x \in \Omega/D \end{cases}$$

As this approach involves the coupling of multiple color layers, this increases the com-

putational complexity. (Duan *et al.*, 2015) solved this problem by designing the Split Bregman (Goldstein and Osher, 2009) algorithm of Eq. (2.6) with the auxiliary variables $\vec{v} = (\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n)$ and Bregman iterative parameters $\vec{b} = (\vec{b}_1, \vec{b}_2, \dots, \vec{b}_n)$. This converts the model into a iterative optimization problem:

$$\begin{aligned} \operatorname{argmin}_{u, \vec{v}} \{ E(u, \vec{v}) = & \sqrt{\sum_{i=1}^n \left(\int_{\Omega} |\vec{v}_i|(x) dx \right)^2} \\ & + \frac{1}{2} \sum_{i=1}^n \int_{\Omega} \lambda_D (u_i - f_i)^2 dx \\ & + \frac{\theta}{2} \sum_{i=1}^n \int_{\Omega} |\vec{v}_i - \nabla_{NL} u_i - \vec{b}_i^{k+1}|^2(x) dx \} \end{aligned} \quad (2.7)$$

such that $\vec{b}_i^{k+1} = \vec{b}_i^k + \nabla_{NL} u_i^k - \vec{v}_i^k$ and $\vec{b}_i^0 = \vec{v}_i^0 = 0$. By optimizing the u and \vec{v} alternatively (i.e. fixing u to optimize \vec{v} and then fixing \vec{v} to optimize u), the solution of Euler-Lagrange equation of u and generalized soft thresholding formula of \vec{v} can be obtained.

$$\lambda_D(u_i - f_i) + \theta \nabla_{NL} \cdot (\vec{v}_i^k - \nabla_{NL} u_i - \vec{b}_i^{k+1}) = 0 \quad (2.8)$$

$$\begin{aligned} \vec{v}_i^{k+1} \approx \max \left(\left| \nabla_{NL} u_i^{k+1} + \vec{b}_i^{k+1} \right| - \frac{\int_{\Omega} |\vec{v}_i^k|(x) dx}{\sqrt{\sum_{i=1}^n \left(\int_{\Omega} |\vec{v}_i^k|(x) dx \right)^2}}, 0 \right) \\ \left(\frac{\nabla_{NL} u_i^{k+1} + \vec{b}_i^{k+1}}{\left| \nabla_{NL} u_i^{k+1} + \vec{b}_i^{k+1} \right|} \right) \end{aligned} \quad (2.9)$$

The Gauss-Seifel iterative scheme is applied to obtain the approximate solution of Eq. (2.8). Auxiliary variable \vec{v} in Eq. (2.9) is the approximate solution. However, it may be corrected by the Bregman iterations and much faster to compute. Figure 2.13 shows the iterative process of the NLCTV inpainting technique.

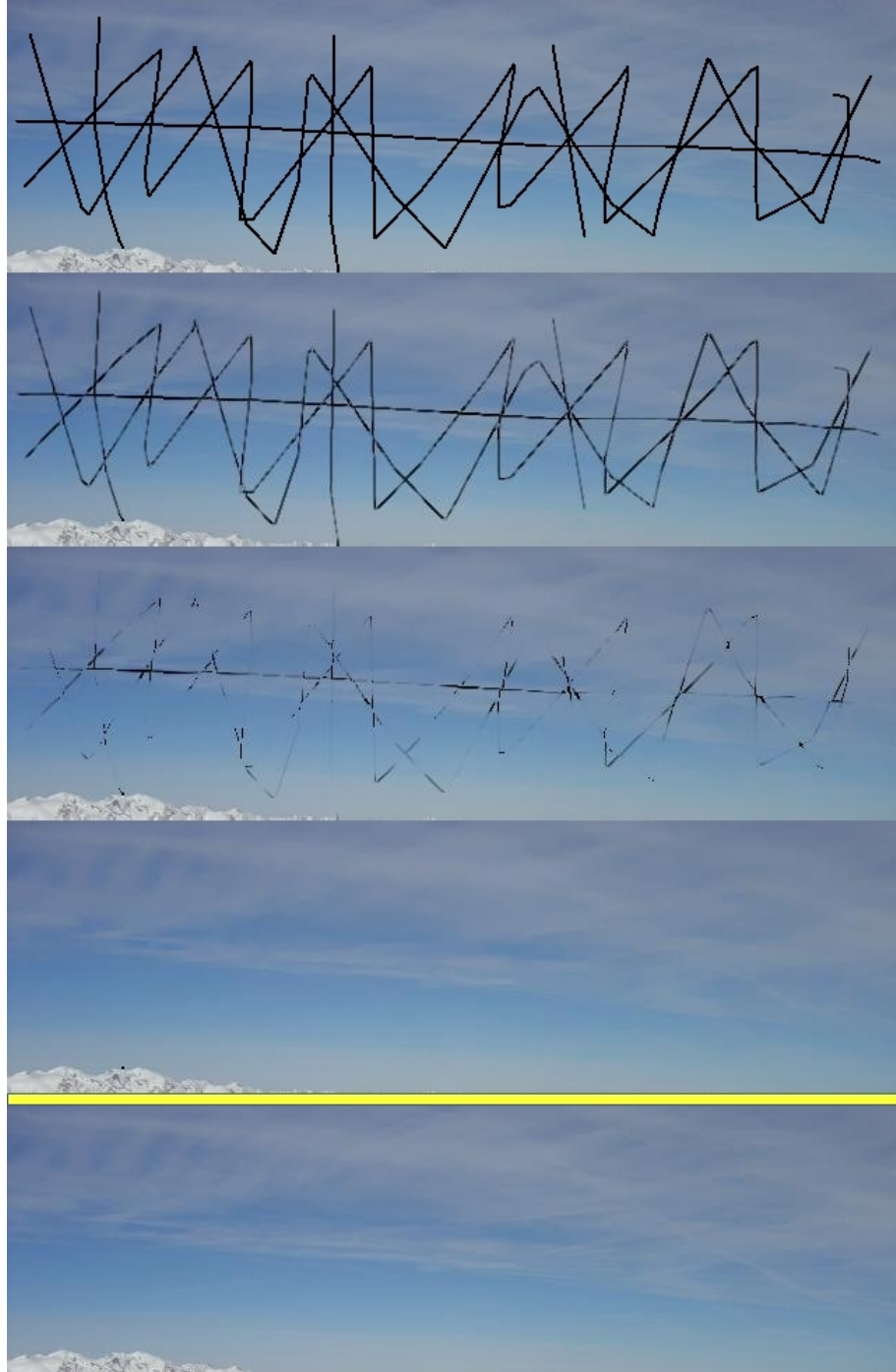


Figure 2.13: These images are the iterative process of NLCTV inpainting approach on a sky image. The image was tainted by irregular dark lines across the whole image (first image). The last image is the original image.

2.3.2 The SFDAE (Stacked Face De-noising Auto Encoders) Deep Learning Model

The Stacked Face De-noising Auto Encoders (SFDAE) Deep Learning model (Pathirage *et al.*, 2015) is another state-of-the-art image reconstruction approach. This model was designed and trained to reconstruct frontal neutral faces from "noisy" faces. The motivation behind the concept of Deep Learning is that many challenges of face-related applications in real-life scenarios consist of non-linear characteristics. Some approaches proposed by Zhu *et al.* (2014) and Zhu *et al.* (2013) are based on Deep Convolutional Networks (DCN) which are capable of learning non-linear feature transformations, such as to reconstruct a frontal face representation from an individual. However, considering the large scale of the framework, it comes with a high demand of large amount of training data to fine-tune the large number of parameters and huge computational power. As a consequence, it limits the range of applications. To overcome these limitations, some approaches based on Deep Auto Encoders (DAE) were proposed (Kan *et al.*, 2014; Vincent *et al.*, 2010; Hinton and Salakhutdinov, 2006). The model's hidden layers allow for an efficient feature learning via orderly non-linear mappings. The simplicity of its design and training process leads to a lower complexity compared to the DCN framework. The SFDAE was designed based on DAE framework inspired by (Kan *et al.*, 2014).

A classic Auto Encoder consists of two main parts: encoders and decoders. Let $f(x)$ be an encoder function which transforms an input vector $x \in \mathbb{R}^d$ into hidden representation $h \in \mathbb{R}^r$ (usually $r < d$):

$$h = f(x) = \Phi(Wx + b) \quad (2.10)$$

where $W \in \mathbb{R}^{r \times d}$ and $b \in \mathbb{R}^d$ are the affine mapping and the bias respectively. $\Phi(x) = \frac{1}{1+e^{-x}}$ (sigmoid) is the activation function which introduces the non-linearity elements into the framework. Decoder is the mapping function $g(h)$ which recovers h into a vector $z \in \mathbb{R}^d$ which is a reconstruction of input vector x :

$$z = g(h) = \Phi(\widehat{W}h + \widehat{b}) \quad (2.11)$$

The SFDAE is a patch-based single-decoder-multiple-encoders framework with an aim to de-noise the contaminated inputs. Supervisory information from the actual frontal neutral faces are used to train and optimize the de-noising layer. As a result, the learnt

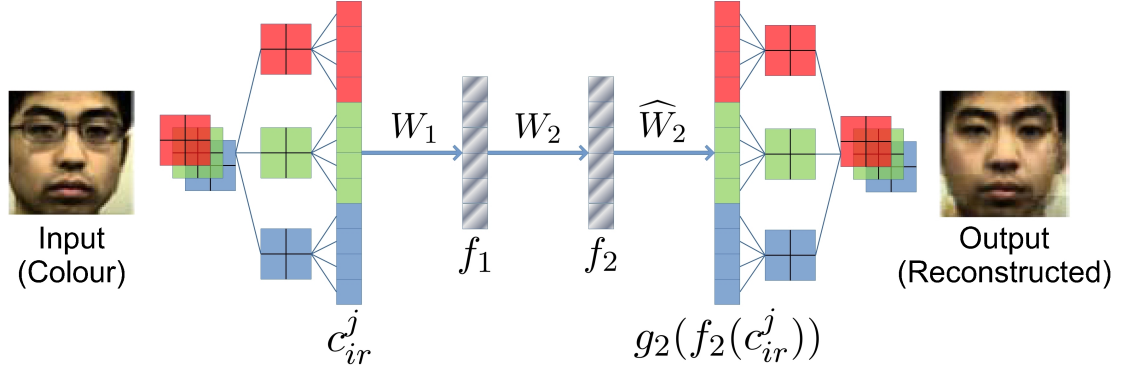


Figure 2.14: This is the overview of SFDAE framework. f_1 represents the low dimensional features derived from the first layer. The features are processed through de-noising process on second layer represented by f_2 . Lastly, the features are decoded to reconstruct the frontal neutral faces.

representation h will contain the highest mutual information between the actual neutral and contaminated faces after de-noising process.

The overall SFDAE framework can be seen in Figure 2.14. The SFDAE consists of three layers. The first layer performs a dimensional reduction with color fusion on the contaminated face input. The second layer de-noises the contaminated segment from the previous layer. The feature space from this layer is the low-dimension discriminative features used for face recognition. The last layer utilizes supervisory information to regularize and optimize the whole learning framework. The cost function of the first layer for optimization of patch j is defined as:

$$\left[W_{l=1}^*, b_{l=1}^*, \widehat{W}_{l=1}^*, \widehat{b}_{l=1}^* \right]_j = \arg \min_{W, b, \widehat{W}, \widehat{b}} \sum_{i=1}^S \sum_{r=1}^{N_i} \left\| c_{ir}^j - g_1(f_1(c_{ir}^j)) \right\|_2^2 \quad (2.12)$$

where S and N_i are the amount of subject identities and amount of images in class i respectively. $f_1()$ and $g_1()$ correspond to encoder and decoder as defined in Eq. (2.10) and Eq. (2.11). c_{ir}^j is the compilation of pixel intensity p_{ir}^j on RGB color channels (Red, Green, Blue) for patch j of image r in class i . In the case of greyscale image, c_{ir}^j only contains one greyscale channel.

$$c_{ir}^j = \begin{cases} \begin{bmatrix} p_{ir}^{jR} & p_{ir}^{jG} & p_{ir}^{jB} \end{bmatrix}^T, & \text{if colour images} \\ \begin{bmatrix} p_{ir}^{jGrey} \end{bmatrix}^T, & \text{if greyscale images} \end{cases}$$

The next step is the de-noising process (second layer) of the low dimensional representation h_{ir}^j derived from the first layer. The cost function is defined as:

$$\left[W_{l=2}^*, b_{l=2}^*, \widehat{W}_{l=2}^*, \widehat{b}_{l=2}^* \right]_j = \arg \min_{W, b, \widehat{W}, \widehat{b}} \sum_{i=1}^S \sum_{r=1}^{N_i} \left\| \left\{ c_{ir}^j \right\}_F - g_2(f_2(h_{ir}^j)) \right\|_2^2 \quad (2.13)$$

where $\left\{ c_{ir}^j \right\}_F$ is the compilation of features of the whole face corresponds to c_{ir}^j . Lastly, the full optimization of the whole system is done by fine-tuning the following equation:

$$\left[W_l^*|_{l=1}^L, b_l^*|_{l=1}^L, \widehat{W}_L^*, \widehat{b}_L^* \right] = \arg \min_{W_l|_{l=1}^L, b_l|_{l=1}^L, \widehat{W}_L, \widehat{b}_L} \sum_{i=1}^S \sum_{r=1}^{N_i} \left\| \left\{ c_{ir}^j \right\}_F - p(c_{ij}^r) \right\|_2^2 \quad (2.14)$$

where $p(x_i) = g_2(f_2(f_1(x_i)))$. Encoder and decoder weights are denoted by $W_l|_{l=1}^L$ and \widehat{W}_L respectively. The representation observed from f_2 (after de-noising) is the input for face recognition.

2.4 Face Recognition Techniques

In chapter 6, our experiments evaluate facial recognition rate as a way to measure the performance of the proposed approaches. Assuming the faces have been detected and aligned, the simplest approach to compare between two faces is by employing the Nearest Neighbour (NN) classifier (Brunelli and Poggio, 1993) where each image is defined from the mixture of its pixel intensity (colour or grey-scale) into a high dimension feature vector in image space. The comparison is conducted by computing and choosing the nearest distance (usually Euclidean distance). Despite its simplicity, the size of the feature vector can grow prohibitively large quickly. For instance, a single 500x500 pixels image will create 250,000 dimension feature vector (750,000 if colour image). This amount leads to expensive computation and large storage/memory requirement. Furthermore, this feature vector is sensitive to noise. For example, variation in illumination might cause multiple face images of the same subject to scatter around in the image space which will drop the recognition rate significantly.

Because of the limitations of the Nearest Neighbour approach, we employ some well-known face classification approaches. In this section, a brief description is provided for two

subspace projection methods PCA (Principal Component Analysis) (Turk and Pentland, 1991) and LDA (Linear Discriminant Analysis) (Belhumeur *et al.*, 1997) and the state-of-the-art SRC (Sparse Representation Classifier) (Wright *et al.*, 2009).

2.4.1 The PCA (Principal Component Analysis)

The PCA (Turk and Pentland, 1991; Belhumeur *et al.*, 1997) is an unsupervised linear dimension reduction approach which projects the feature vector into a lower dimension subspace while maximizing the scatter of the projected data. The intuitive idea here is to preserve a smaller portion of the data while still represents the majority information of the original data. This reduction improves the computation efficiency and removes undesirable noises in the data.

If we consider a set of N face images (input) $\{x_1, x_2, \dots, x_N\}$ represented by a d -dimension feature vector for each image, the aim is to define a mapping W to project all input feature vectors into lower m -dimension ($m < d$) output feature vectors $\{y_1, y_2, \dots, y_N\}$. Each output is calculated as

$$y_i = W^T x_i \text{ for } i = 1, 2, \dots, N \quad (2.15)$$

We calculate the mapping W by maximizing the determinant of total scatter matrix S_t in the following equations:

$$S_t = \sum_{i=1}^N (x_i - \mu)(x_i - \mu)^T \quad (2.16)$$

$$W_{optimal} = \arg \max_W |W^T S_t W| \quad (2.17)$$

where μ is the mean image of all input images. $W_{optimal}$ contains m eigenvectors (also known as Eigenfaces) of S_t with the largest eigenvalues. Other eigenvectors with smaller eigenvalues are usually associated with unwanted noise which can decrease the performance of facial recognition rate and therefore should be discarded.

After learning the optimal mapping $W_{optimal}$, all the gallery images are projected into the new subspace. In the testing stage, the query image is also projected into the corresponding

subspace and the distance to each gallery image is measured. The Nearest Neighbour (NN) classifier is then applied on this newly defined subspace to determine the identity of the query face image.

2.4.2 The LDA (Linear Discriminant Analysis)

The LDA (Belhumeur *et al.*, 1997) is also another subspace projection technique. This is different from the PCA in the context of the objective of projection. The PCA attempts to maximize the data variance in the new subspace, while the LDA maximize the separability/discrimination of the data classes (e.g. face identity). Since the main purpose is to recognize faces, a strong discriminative projection is more desirable intuitively.

The LDA is a supervised approach where it utilizes the identity/class information from the training dataset to train a discriminative classifier. To be more specific, the LDA projects the feature vectors into a new subspace with a requirement that the face images belong to the same subject are clustered together and the clusters of faces on different identities are far away to each other. In summary, the LDA minimizes the intra-class (within-class) distance and maximizes the inter-class (between-class) distance.

The subspace projection is similar to Eq. (2.15). However, the mapping W is now defined based on the inter-class scatter matrix and intra-class scatter matrix. Assuming there are c classes (unique face identities) in the dataset, the between-class scatter matrix (S_b) and within-class scatter matrix (S_w) are defined as:

$$S_b = \sum_{i=1}^c N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (2.18)$$

$$S_w = \sum_{i=1}^c \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T \quad (2.19)$$

where μ_i and N_i are the mean image and amount of images respectively in class X_i . $W_{optimal}$ is then computed by maximizing the ratio between determinants of S_b and S_w :

$$W_{optimal} = \arg \max_W \frac{|W^T S_b W|}{|W^T S_w W|} \quad (2.20)$$

where the solution can be derived from a generalized eigenvectors and eigenvalues formulation:

$$S_b w_i = \lambda_i S_w w_i \quad , \quad i = 1, 2, \dots, m \quad (m \leq c - 1) \quad (2.21)$$

The computation in Eq. (2.21) is only feasible if S_w is a nonsingular (invertible) matrix. Unfortunately, this assumption is most likely not satisfied due to the fact that S_w cannot reach full rank since the number of image samples N is usually much smaller than the number of pixels (d dimension). However, this problem has been solved by incorporating the PCA at the early stage to project the feature vectors into a lower dimension subspace to ensure that S_w is nonsingular. This approach (also known as Fisherfaces) computes $W_{optimal}$ derived from two projections W_{pca} and W_{lda} :

$$W_{optimal}^T = W_{lda}^T W_{pca}^T \quad (2.22)$$

$$W_{pca} = \arg \max_W |W^T S_t W| \quad (2.23)$$

$$W_{lda} = \arg \max_W \frac{|W^T W_{pca}^T S_b W_{pca} W|}{|W^T W_{pca}^T S_w W_{pca} W|} \quad (2.24)$$

The Nearest Neighbour (NN) classifier is also applied on the projected feature vectors in order to recognize the identity of the query face image.

2.4.3 The SRC (Sparse Representation Classifier)

The SRC (Wright *et al.*, 2009) is one of the state-of-the-art approaches to perform classification. This approach has an assumption that the training samples for each identity contains sufficient variations (e.g. facial expressions) spanning the whole face space for a robust facial recognition. However, if some prior knowledge of the query images are known, then less variations of the training samples can be tolerated. For instance, if we know that a query is the mugshot (frontal face without extreme facial expressions), then the SRC will still be able to recognize it even with limited mixture of facial expressions in the training samples. For any query face image, it can be represented by

the linear combination (with coefficients c) of the whole training samples. The intuitive expectation is that coefficient C will be sparse (contain mostly zero-valued elements) with the exception on the training samples of the same identity. This sparse representation will immediately expose the identity of the query face images since it is easy to notice which training subject is dominant in coefficient c .

Assuming we have a sufficiently large set of N d -dimensional training samples $X = \{x_1, x_2, \dots, x_N\} \in \mathbb{R}^{d \times N}$, the expectation is that a query image y can be represented as a linear combination of X and sparse coefficient c :

$$y = Xc \quad , \quad c \in \mathbb{R}^N \quad (2.25)$$

In order to ensure that coefficient c is as sparse as possible while satisfying Eq. (2.25), one needs to solve the following problem:

$$c_0 = \min \|c\|_0 \quad \text{s.t.} \quad y = Xc \quad (2.26)$$

which minimizes the amount of non-zero elements in c through ℓ^0 -norm. Unfortunately, Eq. (2.26) is considered a NP-hard problem which implies that it is difficult to solve it efficiently. However, it has been discovered (Donoho, 2006; Candes *et al.*, 2006; Candes and Tao, 2006; Sharon *et al.*, 2009) that the same solution can be obtained with ℓ^1 -norm with the condition that c is sufficiently sparse:

$$c_1 = \min \|c\|_1 \quad \text{s.t.} \quad y = Xc \quad (2.27)$$

The ideal scenario is that the query image is "clean" (no unwanted noise). However, the real-life scenario will not always satisfy this condition. In order to improve the robustness to noise, (Wright *et al.*, 2009) extend Eq. (2.25) and Eq. (2.27) respectively into:

$$y = Xc + z \quad (2.28)$$

$$c_1 = \min \|c\|_1 \quad \text{s.t.} \quad \|Xc - y\|_2 \leq \varepsilon \quad (2.29)$$

which incorporates noise vector z and noise level ε in the equations to anticipate noise in the image. However, because ε is difficult to predict beforehand, one approach to solve this by employing Lasso (Tibshirani, 1996) with sparsity regularization parameter λ :

$$\min_{c,z} \|y - Xc + z\|_2^2 + \lambda(\|c\|_1 + \|z\|_1) \quad (2.30)$$

After the coefficient c is computed, recognition can be done by choosing the class i which produces the smallest image reconstruction residue on its corresponding coefficients:

$$\min_i r_i(y) = \|y - X\delta_i(c)\|_2 \quad (2.31)$$

where $\delta_i : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is a function to choose only the elements in coefficient c corresponding to class i .

2.5 Databases

All the proposed approaches were tested on various publicly available facial image databases. The usage of all these images is limited to research purpose only. In this section, brief descriptions are provided for each database. However, the detail of experiment configuration (e.g. amount of chosen images, how to define training/testing set) will be described separately on experiment section in each chapter. Since our research scope is on frontal/near-frontal faces, we did not include non-frontal/profile face images in our experiments. The environment of the databases can be divided into two categories: Controlled and Uncontrolled.

2.5.1 Controlled Face Databases

Controlled database is the collection of images captured inside a laboratory/room (indoor) based on some rules/restrictions. The purpose of this strict management is to minimize the possibility of unwanted noise in the data and isolate the problem to be solved. For example, the illumination can be adjusted (e.g. not too bright or too dark) for facial recognition techniques. Some of the restrictions include:

- Illumination (brightness and uniformity)
- Facial Expressions (e.g. neutral, happy, sad)
- Occlusions (e.g. sunglasses, scarf)
- Pose (angle of the face with respect to camera)

2.5.1.1 AR Dataset

There are over 3000 colour face images in AR database (Martínez and Benavente, 1998) (Martínez, 1998) captured from 136 people (76 males and 60 females). However, only photographs from 116 people (63 males and 53 females) were obtained properly on all sessions. Each participant was required to attend two sessions (2 weeks apart). Although all the images are only frontal faces, it has 13 variations on facial expressions (neutral, smile, anger, and scream), illumination (lighting from left, right, and both), and occlusions (sunglasses and scarf). (Figure 2.15).

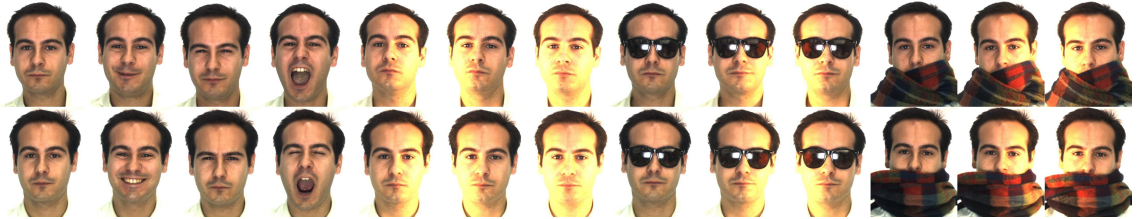


Figure 2.15: 13 image variations on 2 sessions for each participant on AR database.

The ground truth of landmarks (Figure 2.16) are available for 112 people provided by Ding and Martinez (Ding and Martinez, 2010) on all facial expressions (neutral, smile, anger, scream). There are 130 landmarks covering all facial components (eyebrows, eyes, nose, mouth, and jawline) which provide decent amount of information to define geometric features. The ground truth of landmarks plays a significant role on our proposed approach.

2.5.1.2 CMU multiPIE

CMU multiPIE (Gross *et al.*, 2010) (Gross, 2010) is a massive face database extended from the Pose, Illumination, and Expression (PIE) database (Sim *et al.*, 2002). It contains more than 750,000 face images from 337 participants with a variation on 6 facial expressions



Figure 2.16: Landmarks ground truth on AR database.

(neutral, smile, surprise, squint, disgust, scream), 15 camera viewpoints (pose), and 19 illumination conditions in 4 sessions over the span of 5 months (Figure 2.17). Frontal faces are recorded on high resolution, thus is suitable for face-related applications (e.g. detection or recognition).



Figure 2.17: Examples of all the facial poses and expressions on CMU multiPIE database.

2.5.1.3 PUT

PUT database (Kasinski *et al.*, 2008) (Schmidt, 2008) emphasizes heavily on the mixture of face poses. 9,971 face images were captured from 100 participants. It means that approximately 100 photographs were obtained from each subject. These 100 images were divided into 5 subsets of image sequences (Figure 2.18). The first 4 subsets are the sequence of faces rotating in different directions on various perspectives. The last subset is the sequence of images without any constraint on expression or poses.



Figure 2.18: Sample images from PUT database. Every 3 images is one subset.

2.5.1.4 FEI

FEI (OLIVEIRA JR and Thomaz, 2006) (Thomaz, 2006) is Brazilian face database from Artificial Intelligence Laboratory of FEI in São Bernardo do Campo, São Paulo, Brazil. It involved 200 participants (100 male and 100 female) with 14 images each. The variations

in the images include profile rotation (from 90° facing right to 90° facing left), slight facial expressions (neutral and smile), and illumination condition. (Figure 2.19)

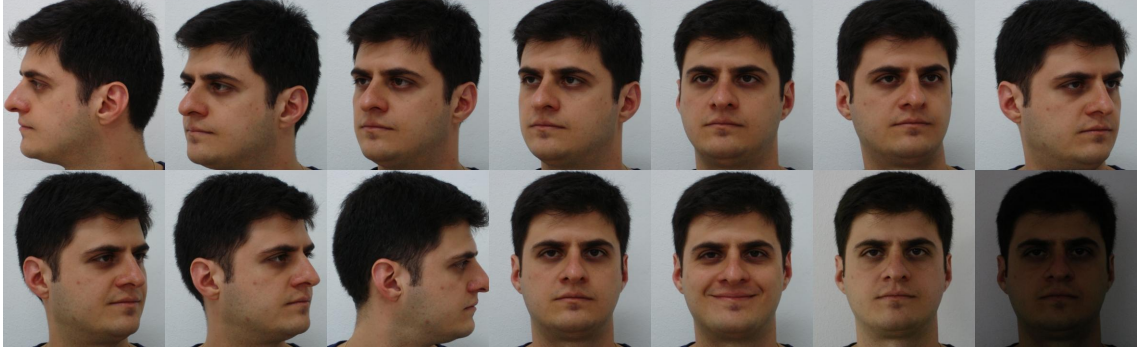


Figure 2.19: Sample images from FEI database.

2.5.1.5 BU-4D

BU-4D (Yin *et al.*, 2008) (Yin, 2008) is mainly for facial expression (anger, disgust, happiness, fear, sadness, surprise). The images of all participants are available in both 3D (point cloud + texture) and 2D (digital image). This database is the extension of BU-3D database (Yin *et al.*, 2006) where a video (approximately 100 frames) is captured for each facial expressions to create a dynamic 3D space of the data. There are 101 participants (43 male and 58 female). We only emphasize on 2D images for our experiments by manually choosing the neutral frontal faces.

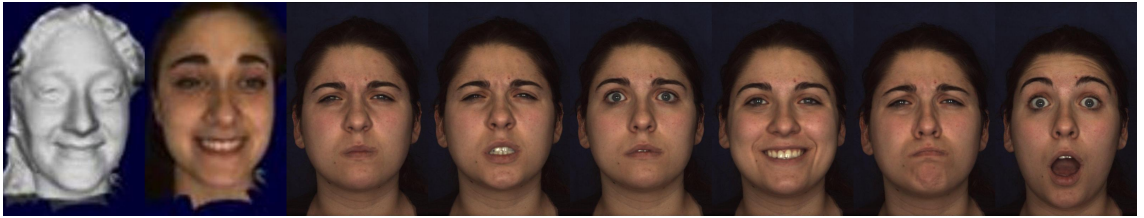


Figure 2.20: 3D and 2D data from BU-4D database.

2.5.1.6 CurtinFaces

CurtinFaces (Li *et al.*, 2013) (Mian, 2013) has both 2D colour images and "depth" information which provides basic information of 3D geometric features of the faces. A standard digital camera (Lumix-DMC-FT1) (high resolution 4000x3000 colour images) and a Kinect sensor (Microsoft) (640x480 colour + depth images) were used in the photography session. Various facial expressions, illuminations, poses, and occlusions were captured from

52 participants along with some combinations (e.g. expression + pose, expression + illumination) leading to 97 images per subject. Some of the examples can be seen in Figure 2.21.



Figure 2.21: range data (depth) and 2D images from CurtinFaces database.

2.5.1.7 CAS-PEAL-R1

CAS-PEAL-R1 (Gao *et al.*, 2008) (Shan, 2008) is another massive database collected under the sponsor of the Chinese National Hi-Tech Program and ISVISION Tech. Co. Ltd. The variations in this database are enormous, particularly in Pose, Expression, Accessories, and Lighting (PEAL). The whole database consists of 99,594 photographs from 1,040 participants taken from 9 camera angles, 5 facial expressions (closed eyes, frown, open mouth, smile, surprise), 6 accessories (3 hat, 3 glasses), and 15 illumination directions. However, only a partial of this dataset is made available to public which is called CAS-PEAL-R1 containing 30,900 grey-scale photographs with less variations. (Figure 2.22)

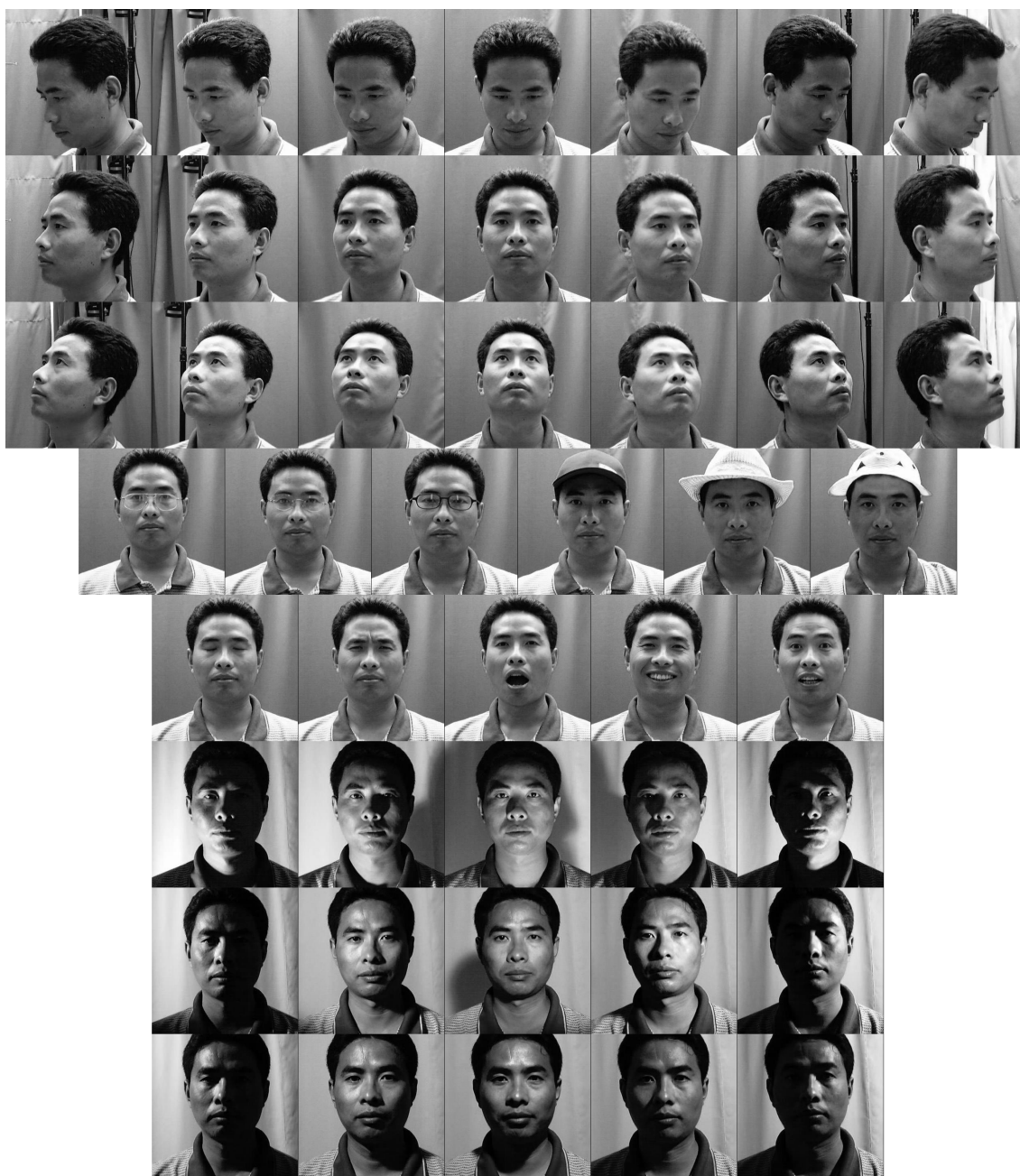


Figure 2.22: Images with variety on (from top) pose, accessories, expression, and illumination from CAS-PEAL-R1.

2.5.2 Uncontrolled Face Database

Uncontrolled database is the opposite of the controlled database where there is no specific restrictions. All photographs could be captured either indoor or outdoor. It is also possible to have multiple faces in a single image. This configuration simulates the real-life scenarios, thus producing a more challenging problem to solve.

2.5.2.1 FDDB

Face Detection Data set and Benchmark (FDDB) (Jain and Learned-Miller, 2010) (Chowdhury, 2010) is a collection of face images obtained from the Faces in the Wild data set (Berg *et al.*, 2004). The primary purpose of this database is for the performance evaluation of face detection techniques. This contains 5171 faces from 2845 images (Figure 2.23). The ground truth of the location of the faces are provided along with the source code to measure the ROC (Receiver Operating Characteristic) curve by plotting the relation between True Positive and False Positive rate.



Figure 2.23: Some collections of face images from FDDB database.

2.5.2.2 AFLW

Annotated Facial Landmarks in the Wild (AFLW) (Koestinger *et al.*, 2011a) (Koestinger *et al.*, 2011b) is a massive-scale real-world collection of face images gathered from an online photo management Flickr. Approximately 25,000 faces are available for the evaluation on automatic facial detection/landmarking and pose estimation (Figure 2.24). There is no particular restrictions on the facial expressions, number of faces in a single image and poses.

2.5.3 INRIAperson

INRIAperson (Dalal and Triggs, 2005) (Dalal, 2005) is not a facial database. It contains 1805 images of human standing on variety of orientation and background. This database is suitable to evaluate the performance of person detection techniques. Our main focus is on the negative training subset. This subset contains 1218 non-person photograph which is suitable to train the face detection system to distinguish between faces and non-faces images. (Figure 2.25)

2.6 Summary

This chapter provides some basic introductions on preliminary knowledge related to the topic of this thesis. First, the basic concept of the pictorial-tree-structured face models by Zhu and Ramanan (2012a) is described along with the availability of open source code (Zhu and Ramanan, 2012b). This explanation gives an intuitive idea on how the model works to extract facial landmarks. Second, we describe the well-known Viola and Jones face detector as it is employed in our proposed facial landmarking system. Third, two image reconstruction approaches are explained briefly where they can be implemented to remove the "noise" from faces. Fourth, three face classification approaches are discussed. Lastly, we provide brief descriptions on all facial/non-facial databases involved in all of our experiments.

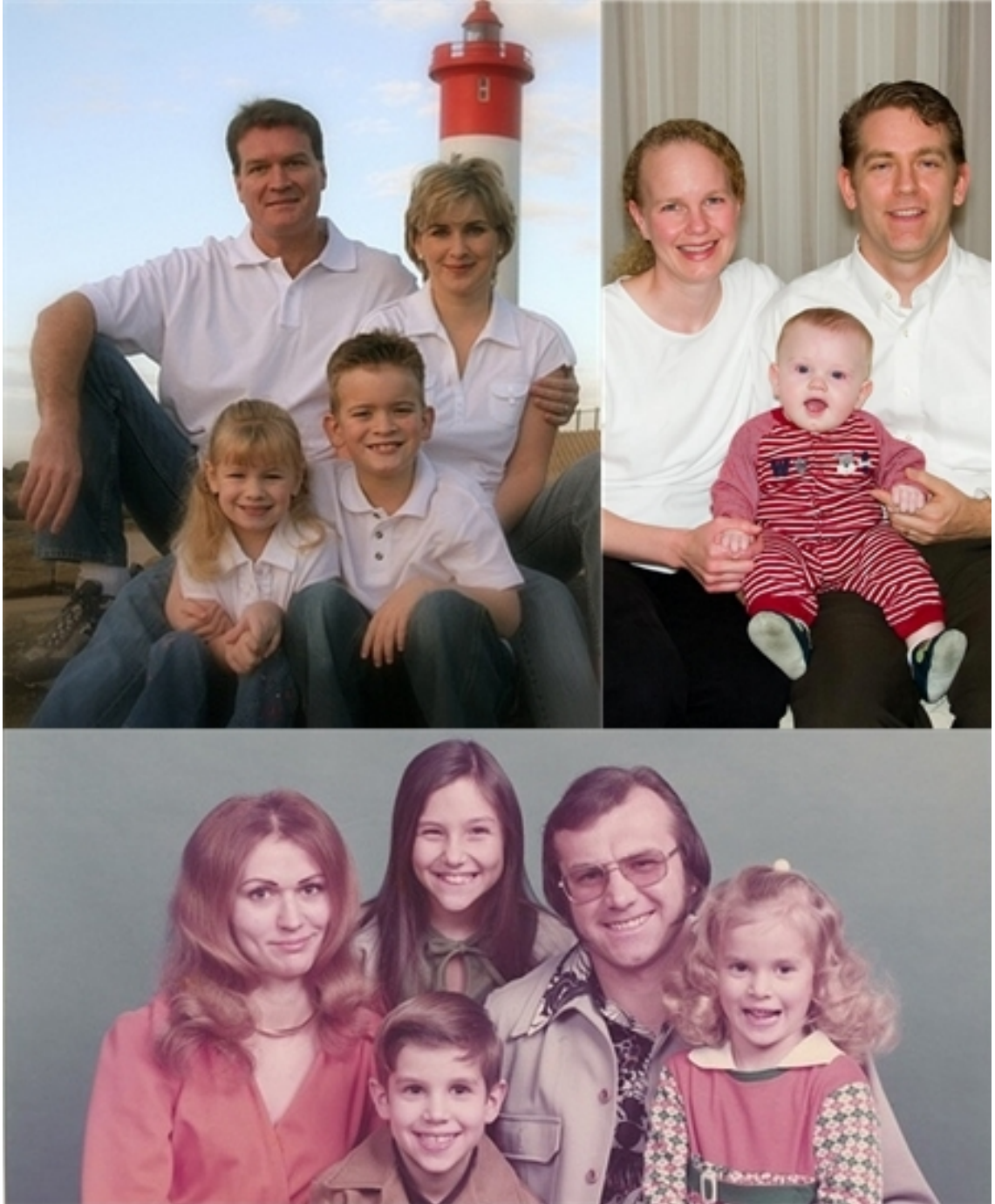


Figure 2.24: Samples of images from AFLW database.



Figure 2.25: Samples of non-facial images from INRIAperson database.

Chapter 3

Facial Landmarks Detector with High Density Landmarks

Face image retrieval based on semantic features extraction requires the description (e.g. shape and size) of facial components (e.g. eyes and mouth). One of the recent research on semantic-based face retrieval is done by Conilione and Wang (2012). Their approach strongly relies on accuracy of the location of facial landmarks to extract geometric information. However, the facial landmarks were acquired manually there by human. This is a time-consuming and impractical task, especially for enormous amount of faces in a large dataset. This is why we need an automatic approach which can detect facial landmarks. Some approaches have been developed to address this particular task such as Zhu and Ramanan (2012a), Le *et al.* (2012), and Valstar *et al.* (2010). The one which caught our attention and considered as the state-of-the-art at the time as claimed by Çeliktutan *et al.* (2013) was developed by Zhu and Ramanan (2012a). As discussed in section 2.1, the concept of tree pictorial structure combined with Histogram of Oriented Gradients (HOG) features produces robust facial landmarking models which make it highly tolerant against face deformations with capability to handle a large variety of faces.

However, Zhu and Ramanan’s face models have a shortcoming of having insufficient number of facial landmarks. In fact, this is a common shortcoming for all existing facial landmark detectors. As explained in section 2.1.4, their frontal face models only provide 68 landmarks. Figure 3.1 shows the sample of facial landmarks extracted by one of their face models. It can be seen that some face regions such as eyes are covered only by six landmarks in each eye. As our objective is to extract semantic features, it is difficult to get an accurate semantic representation with such few landmarks. For example, 6 landmarks on the eye can only describe limited shapes such as hexagon or trapezoid which is not an accurate depiction of eye silhouettes. The motivation of these existing landmark detectors is to detect the existence and positions of facial components without consideration of semantic descriptions of them. For this reason, we are motivated to propose a more sophisticated and accurate frontal face model with an aim to use more landmarks for our purpose of semantic description. For this purpose, we redesigned the architecture of Zhu and Ramanan’s face models to have more landmarks in coverage of facial components

especially on eyebrows, eyes and nose. The aim is to have a possible semantic description of facial components with these added landmarks.

For this purpose, the AR database (Martínez and Benavente (1998)) was selected to train the frontal face model on various expressions because of the availability of high density facial landmarks in ground truth (Ding and Martinez (2010)). For the rest of this chapter, the proposed face model is referred as the **AR model**. Figure 3.2 shows the examples of facial landmarks detection testing to see the significant difference of the information provided by the AR model. One can see clearly that shape description becomes possible with the proposed **AR model**.

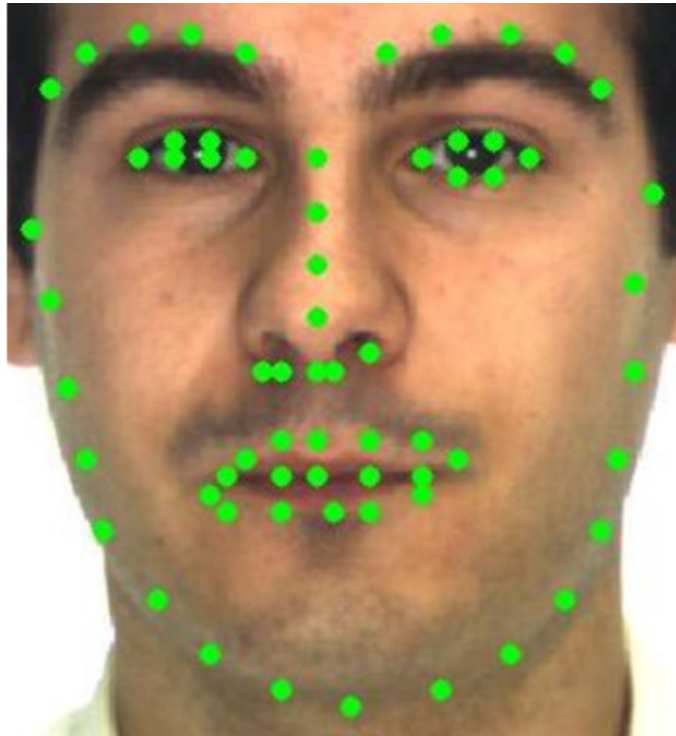


Figure 3.1: Facial Landmarks from the Independent-1050 model is not sufficient to extract semantic features. (© 2014, IEEE)

In terms of performance evaluation for different landmark detectors, the accuracy of the location of important facial landmarks is used. We will compare the performance with Zhu and Ramanan's face model Independent-1050 which has been claimed to perform with the highest accuracy among all their proposed models. Furthermore, we also compare with another robust facial landmarking approach proposed by Le *et al.* (2012). They developed a model called the CompASM which is an improvement of the classic statistical face model Active Shape Model (ASM) (Cootes *et al.* (1995)). Instead of imposing the face model into a holistic individual Gaussian model, Le *et al.* developed a component-based ASM

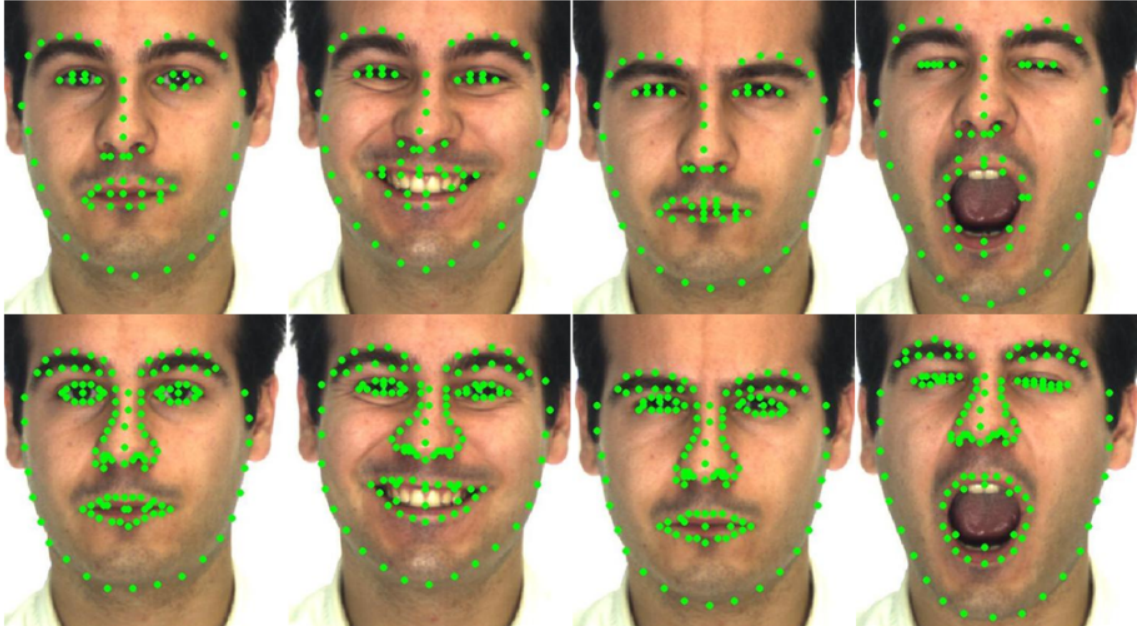


Figure 3.2: Examples of the facial landmarking results between Independent-1050 models (**top**) and our proposed model (**bottom**). (© 2014, IEEE)

model to make it more robust against occlusions and various face expressions including irregular one such as winking. Their work was inspired by the component-based approach in Huang *et al.* (2007) and pictorial structure by Felzenszwalb and Huttenlocher (2005).

Lastly, many studies have shown that various colour spaces can significantly improve the accuracy of face recognition (Yang and Liu (2008)). However, to the best of our knowledge, not many researches have conducted on the impact of colour spaces on facial landmarks detection. This motivates us to pursue this investigation with the AR model. We did the face model training and testing accordingly on three colour spaces i.e. grey-scale, HSV, and RGB-NII Yang *et al.* (2010) and analyze the effects on the performance.

The structure of this chapter is organized as follows. Section 3.1 describes the tree structure when adding more landmarks. Section 3.2 addresses the training setup and dataset used for training AR model. All experimental setup and protocols are provided in section 3.3 along with the results. The contributions of this chapter are summarized in section 3.4.

3.1 Model Creation

In Milborrow and Nicolls (2008), they claimed that a face model with a high level of facial landmarks density is more likely to have a better semantic description. Since our purpose is to have an accurate face model with more landmarks, we need to train the proposed AR model. For such aim, we would rearranged the architecture of Zhu and Ramanan’s Independent-1050 frontal face models when adding more landmarks. Figure 3.3 visualizes the changes we applied. The majority of the modifications were focused on the eyebrows, eyes, and nose to cover these facial components with a more accurate contour instead of a simple curve. Therefore, extracting semantic features will be more feasible in future. Despite the significant changes on the model architecture, we still maintain the property of a tree structure (no connected loop) for global optimal solution when we use dynamic programming to seek the solution as stated by Zhu and Ramanan. Furthermore, the symmetry of facial components (landmarks) and the relative positions between them are also preserved.

3.2 Model Training and AR Database

The training face images for the AR model are chosen from AR database (Martínez and Benavente (1998)). Our reason to choose this database is due to the availability of dense landmarks in ground truth (Ding and Martinez (2010)) as described in section 2.5.1.1. This amount is much higher than the Independent-1050 models trained on CMU multiPIE database (Gross *et al.* (2010)) which only provides 68 landmarks on frontal faces. Since the face images on AR dataset are divided into two sessions (2 weeks apart between sessions), we select the first session for training and the other for testing. The face images are chosen based on two categories. First, we avoid various illumination and occlusion (sunglasses and scarf) scenarios, thus leaving us with four facial expressions (neutral, smile, anger, scream). Second, as the landmarks in ground truth are only available for limited participants, the proposed face model is only trained on 112 distinct individuals (58 men, 54 women).

The AR model is trained with the same training method provided by Zhu and Ramanan (2012b) with the spatial resolution variable for HOG cells set as 4. 448 face images (112 per facial expression) were used as input for positive image set. On the other hand, 1218 non-face images from INRIAperson database (Dalal and Triggs (2005)) were used as negative image set. As the Independent-1050 models gains the highest accuracy due to its non-sharing-parts trait, we also develop the AR model with independent landmarks

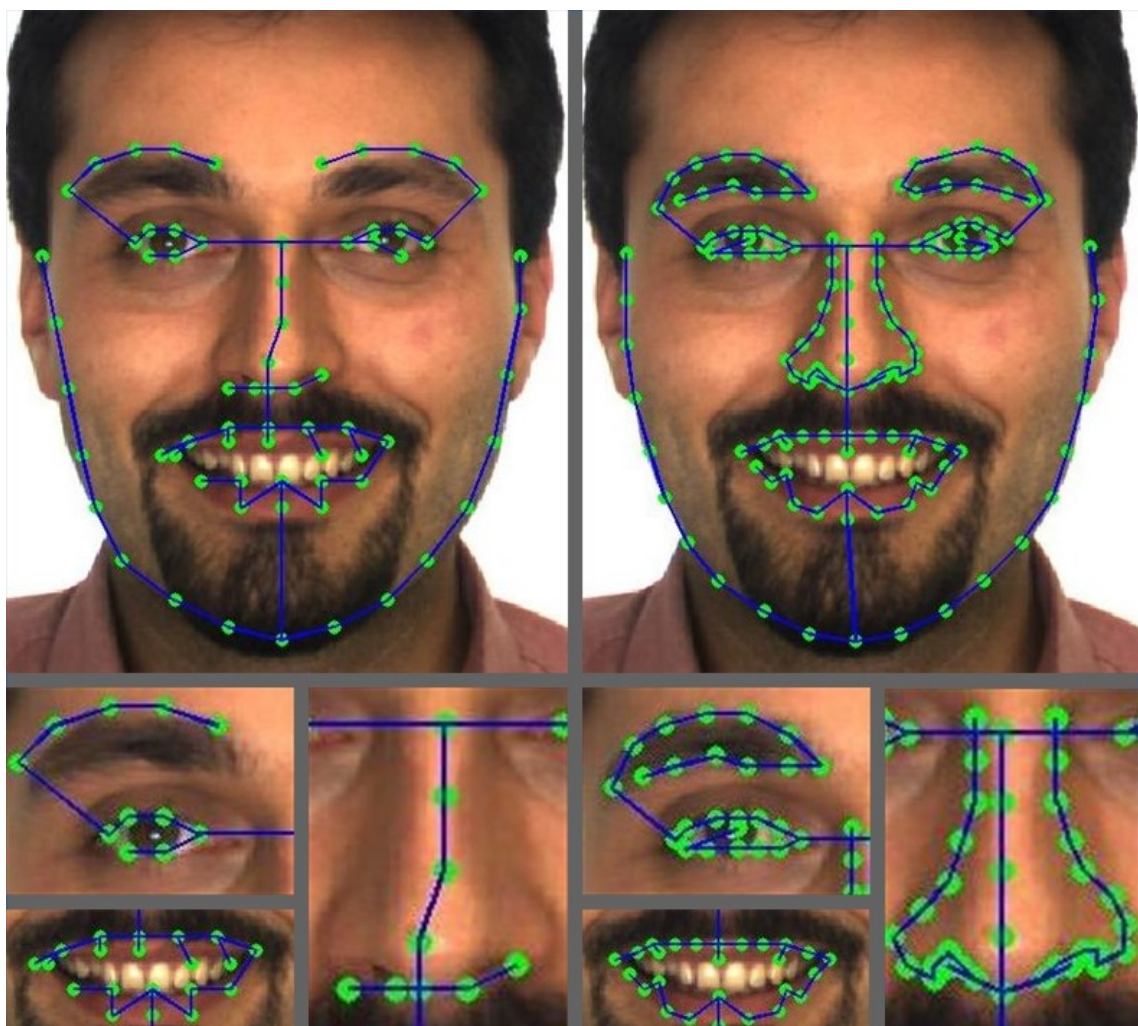


Figure 3.3: The tree structure of Independent-1050 model (**left**) and proposed AR model (**right**). The tree restructuring was made to depict better geometric descriptions and improve the accuracy rate. (© 2014, IEEE)

(130 unique landmark descriptions for each facial expression). The visualization of both Independent-1050 and AR model can be observed in Figure 3.4. It can be clearly seen that the proposed AR model can provide better information on face description.

3.3 Experiments

In order to evaluate the performance of the proposed AR model, we have conducted three separate experiments. First, we compare it with the Independent-1050 models to measure the performance improvement on both accuracy and geometric descriptions. Second, we compare with another robust approach CompASM proposed by Le *et al.* (2012) with the same evaluation metrics. Lastly, the AR model is integrated on various color spaces on both training and testing for performance comparison. All experiments are done on 112 subjects (along with the landmark ground truth) in the second session of AR database.

3.3.1 Evaluation Protocols

We employ standard procedures of evaluating performance of facial landmarks detector as mentioned by Çeliktutan *et al.* (2013). The first procedure is to compare the average difference/distance between the detected landmarks and the ground truth. This distance may also be referred as *relative error*. In order to produce a consistent result regardless the size of the faces, the error rate is normalized by dividing with the corresponding Inter-Ocular Distance (IOD) (distance between both eye centres). Let G and Q be the sets of face images with ground truth and testing query face images respectively. The calculation can be defined as:

$$\text{dist}_{\text{IOD}}(G_i^j, Q_i^j) = \frac{\|G_i^j - Q_i^j\|_2}{\text{IOD}_i} \quad (3.1)$$

where the error rate is measured by calculating the Euclidean distance between landmarks j on face image i , then normalized with the corresponding IOD. Assuming there are N face images and L landmarks to be tested on each face, the average relative error can be computed as:

$$\text{Relative Error} = \frac{\sum_{j=1}^L \sum_{i=1}^N \text{dist}_{\text{IOD}}(G_i^j, Q_i^j)}{L * N} \quad (3.2)$$

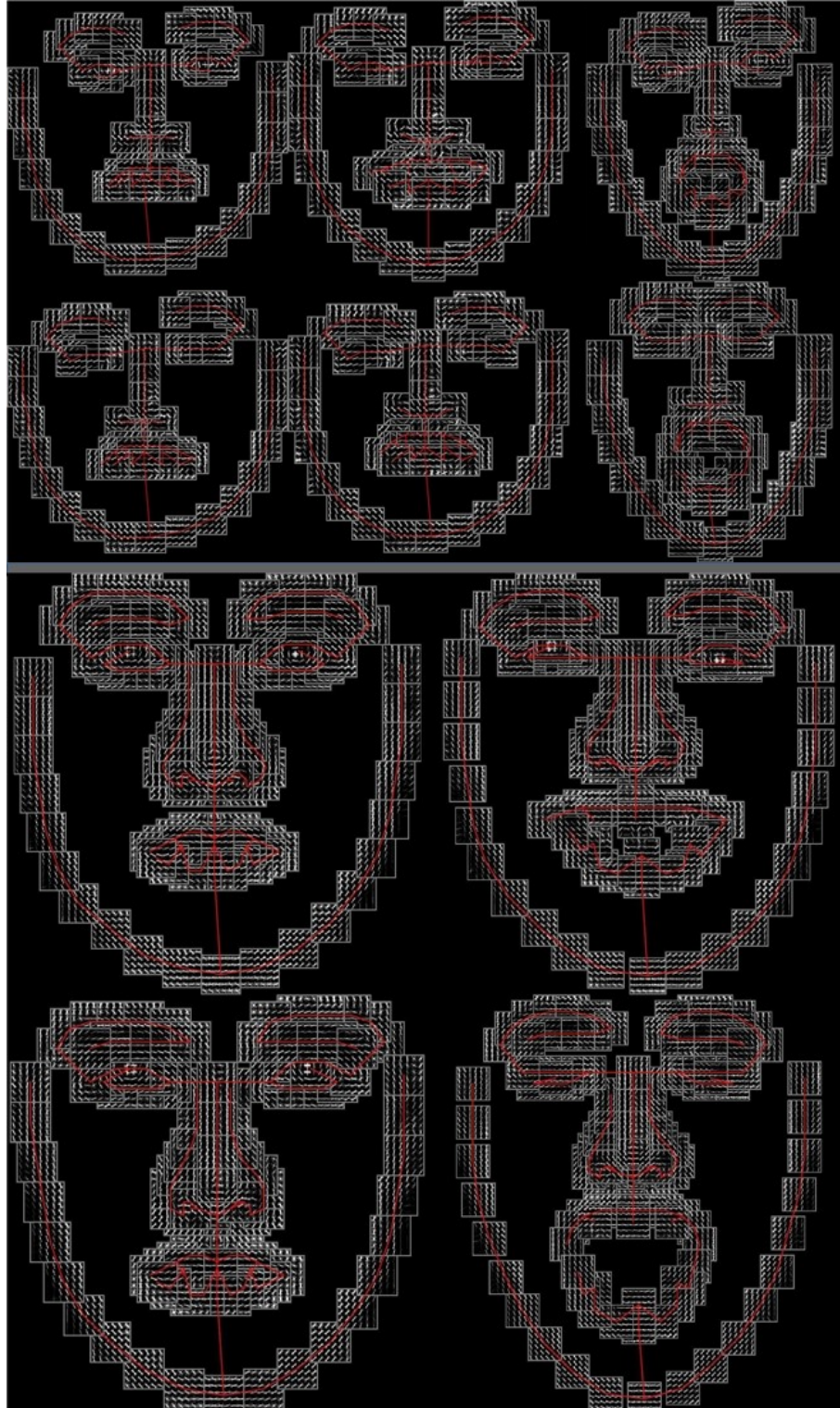


Figure 3.4: Visualization of frontal face models with various facial expressions from Independent-1050 (**top**) and AR model (**bottom**). (© 2014, IEEE)

Even though the relative error is able to show decent indication of the facial landmarking performance, it is sensitive to landmark outliers which can increase the error rate significantly. To further improve the validity of the result, we employ another procedure. The second procedure is to measure the *detection rate* of the extracted landmarks. It counts the number of landmarks which are detected in a reasonable distance to the landmark ground truth. In other words, the detected landmarks have to be inside a range of particular thresholds. In this scenario, the thresholds are defined based on the percentage of the corresponding IOD. We measured on three categories: 5%, 10%, and 20% as shown in Figure 3.5.

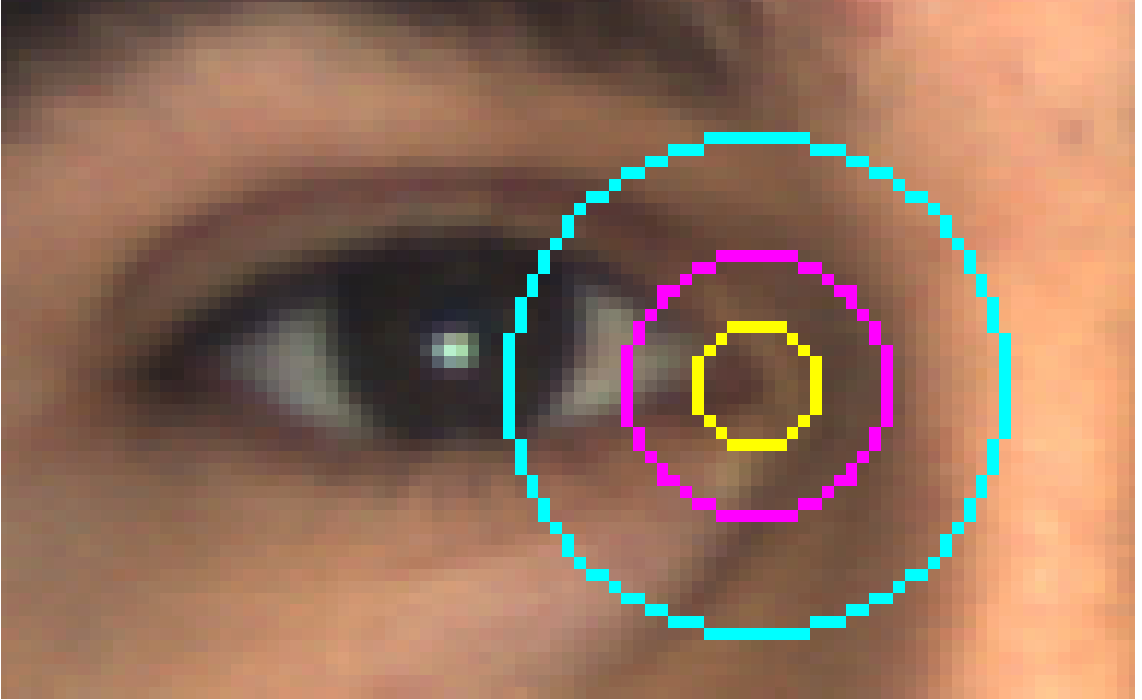


Figure 3.5: 3 different thresholds for detection rate on eye corner. Starting from the smallest circles are the 5%, 10%, and 20% of IOD respectively. (© 2014, IEEE)

As there is a significant difference on the amount and location of landmarks between Independent-1050, CompASM, and AR model, we need to choose which landmarks to be used for fair comparison. We decided to compare 17 fiducial landmarks from the *m17* set (Çeliktutan *et al.* (2013)). These landmarks include eyebrow corners (4 landmarks), eye corners and centres (6 landmarks), nose tip and sides (3 landmarks), and surrounding the mouth (4 landmarks). These landmarks are chosen because its presence is consistent among all the face models to compare (except for 2 landmarks on side nose for Independent-1050). Furthermore, these landmarks are considered useful and important because of its stability and reliability for facial recognition/tracking applications on various facial expressions. Figure 3.6 visualizes the landmarks set L for Independent-1050, CompASM,

and AR model.

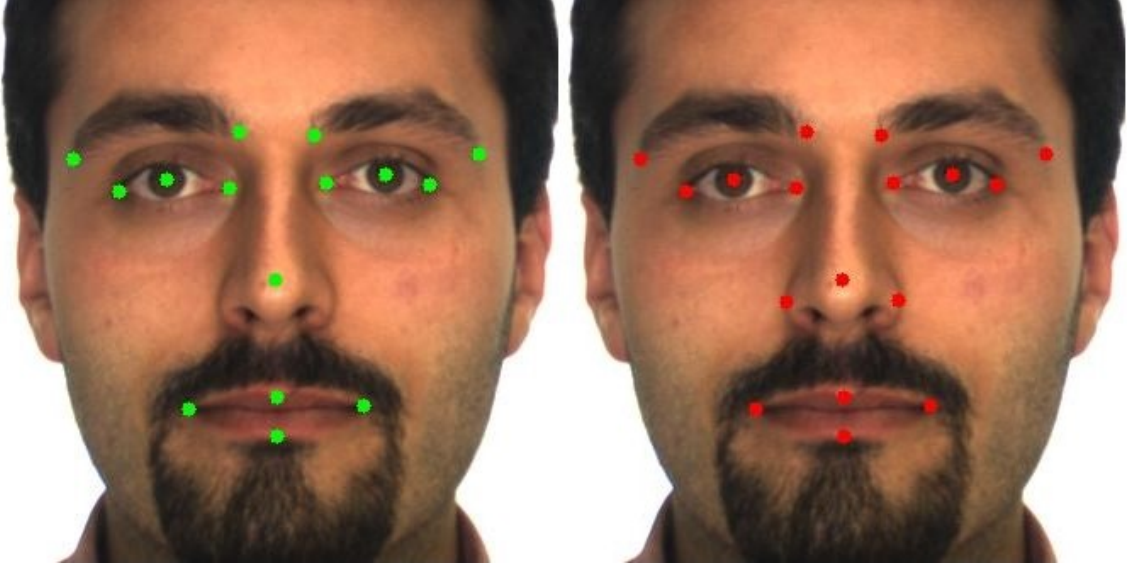


Figure 3.6: 15 landmarks for comparison on Independent-1050 (**left**) and 17 landmarks for comparison on CompASM and AR model (**right**).

Except for these two standard procedures, we also evaluate the accuracy of geometric description of the landmarks on three facial components. This is new as previous landmark detectors mainly concern about component existence and locations. This is done by measuring the width and height of facial components $K = \{left\ eye, right\ eye, mouth\}$ and compare it with the ground truth. The error rate is also normalized by the corresponding IOD. The width is calculated as the distance between the leftmost and rightmost of landmarks on that corresponding facial component (horizontal). The height is also computed in the same manner except that the direction is vertical. Figure 3.7 visualizes the measurement of both width and height. Using the same image sets G and Q containing N face images, the width error rate for each facial component in K can be computed as:

$$\text{width}_{\text{IOD}}(G_i^k, Q_i^k) = \frac{\text{width}_{\text{diff}}(G_i^k, Q_i^k)}{\text{IOD}_i} \quad (3.3)$$

$$\text{Width Error Rate}^k = \frac{\sum_{i=1}^N \text{width}_{\text{IOD}}(G_i^k, Q_i^k)}{N} \quad (3.4)$$

While the height error rate can be defined as:

$$\text{height}_{\text{IOD}}(G_i^k, Q_i^k) = \frac{\text{height}_{\text{diff}}(G_i^k, Q_i^k)}{\text{IOD}_i} \quad (3.5)$$

$$\text{Height Error Rate}^k = \frac{\sum_{i=1}^N \text{height}_{\text{IOD}}(G_i^k, Q_i^k)}{N} \quad (3.6)$$

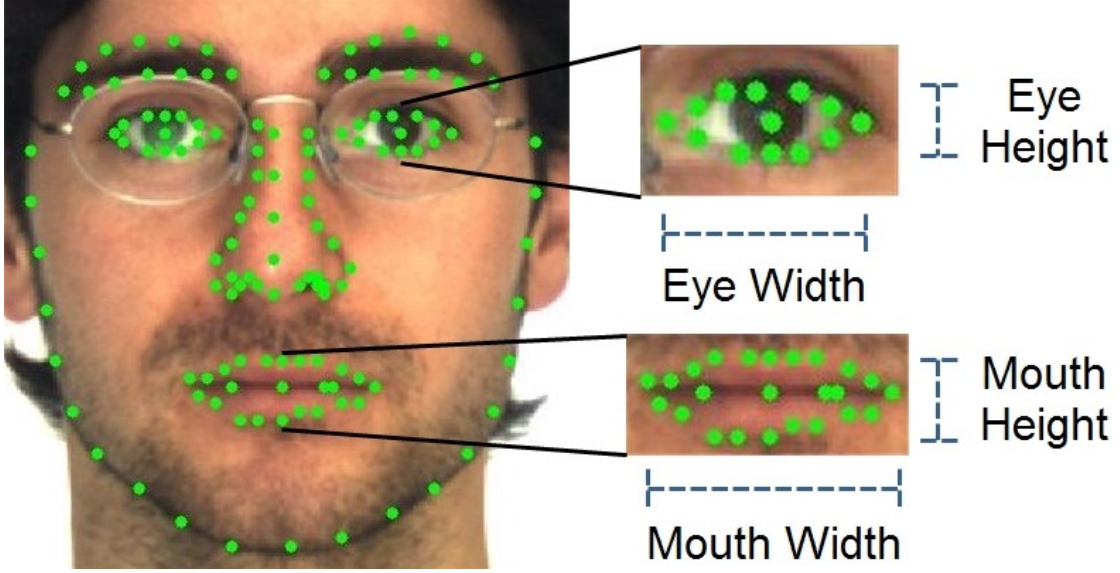


Figure 3.7: Width and height of the facial components are calculated as the largest distance between landmarks on horizontal (**x-axis**) and vertical (**y-axis**) directions respectively.

As for the third experiment, we conduct an investigation on the impact of various colour spaces on the AR model in detecting facial landmarks (Tkalcic and Tasic (2003) provides some preliminary descriptions of some early colour spaces). This experiment is motivated by the fact that colour information provides essential features for improving the performance of face image retrieval or recognition (Yang and Liu (2008)). However, as far as we know, the studies on facial landmarks detection are not found.



Figure 3.8: Four colour spaces: RGB, grey-scale, HSV, and RGB-NIL.

In this case, the AR model is trained in RGB (Red, Green, Blue) colour space which is the commonly used colour for digital images. We conduct the experiment with three other colour spaces: grey-scale, HSV (Hue, Saturation, Value), and RGB-NII as proposed by Yang *et al.* (2010). The visualization of these colour spaces can be seen in Figure 3.8. RGB-NII is defined as the normalized RGB with across-color-component colour space normalization technique (CSN-II) as follows:

$$\begin{bmatrix} \tilde{R}_{II} \\ \tilde{G}_{II} \\ \tilde{B}_{II} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5774 & 0.7887 & -0.2113 \\ -0.5774 & -0.2113 & 0.7887 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

The face images for training and testing were converted to each colour space. We then trained 4 variation of AR models based on the corresponding colour space. Finally, we measure the relative error and detection rate of all 130 landmarks compared to the ground truth. In summary, we conducted performance evaluations by measuring:

- Relative error of 17 landmarks on *m17* set (eyebrow corners (4 landmarks), eye corners and centres (6 landmarks), nose tip and sides (3 landmarks), and surrounding the mouth (4 landmarks)). (15 landmarks only for Independent-1050 excluding 2 landmarks on side nose)
- Detection rate with thresholds 5%, 10%, and 20% of IOD on landmarks set *m17*.
- Width and height error rate for eyes and mouth.
- Relative error and detection rate of AR model on 4 different colour spaces on all 130 landmarks.

3.3.2 The Independent-1050 Model VS the AR Model

We first compare the performance of the proposed AR detector and the Independent-1050 model. The first part of the result on relative error and detection rate is shown in Table 3.1. It can be seen that the Independent-1050 models produce approximately twice the error rate compared to our proposed AR model. Similar improvement can also be observed by the detection rate. the AR model achieves much higher detection rate even in the smallest threshold with **37.19%** improvement. These results show that our proposed model can

detect important landmarks more accurately. One can find that adding more landmarks can improve the detection rate significantly.

Table 3.1: Relative error and detection rate from Independent-1050 and AR model. (© 2014, IEEE)

Model	Relative Error	5% IOD	10% IOD	20% IOD
Independent-1050	0.0726	41.29%	77.26%	96.73%
AR model	0.0365	78.48%	96.35%	99.75%

The second part of the result on geometric shape is shown in Table 3.2. Once again, the proposed AR model can outperform the Independent-1050 significantly, especially on the eyes by large margin. This is to be expected since some of the detected landmarks do not cover the eyes properly in the Independent-1050 model. Some of the examples can be seen in Figure 3.9. These experiments show that the proposed model is much better than the Independent-1050 model.

Table 3.2: Width and height error rate from Independent-1050 and AR model. (© 2014, IEEE)

Component	Model	Width Error	Height Error
Right Eye	Independent-1050	0.0956	0.0561
	AR model	0.0267	0.0233
Left Eye	Independent-1050	0.0941	0.0515
	AR model	0.0254	0.0214
Mouth	Independent-1050	0.0438	0.0522
	AR model	0.0361	0.0403

3.3.3 The CompASM model VS AR Model

In this section, we will compare the proposed AR Model with another well-known detector, the CompASM model. The necessity is that the CompASM model is totally created from different perspectives and it does not belong to the same category of the proposed AR Model and the Independent-1050 model. The relative error and detection rate are shown in Table 3.3 and one can see that the proposed model is much better than the CompASM model. Furthermore, it turns out that some faces can not be detected by the CompASM model due to a low fitting score and we will discuss this issue in next chapter. There are 11 false negative cases i.e. 1 smiling face and 10 scream expressions. We avoid involving these face images for a fair comparison.

The result on geometric description is listed in Table 3.4. This also shows a significant



Figure 3.9: Some testing result from Independent-1050 (**left**) and AR model (**middle**). The landmarks in ground truth are shown in the last column (**right**).

performance gap dominated by our proposed AR model. The enormous gap on width and height error rates on mouth drew our attention. After conducting intense observation, we discovered that the CompASM model does not work well with the scream facial expression. The examples can be seen in Figure 3.10. As this affects the performance significantly, we repeat the experiments after ignoring scream expression. The revised result is shown in Table 3.5. One can find that the proposed AR model still outperforms the CompASM model significantly and quite consistent on the other three facial expressions. However, the CompASM model did quite well on extracting width and height of the eyes as the error rate on a par with AR model and even slightly better on neutral expression.

Table 3.3: Relative error and detection rate from CompASM and AR model. (© 2014, IEEE)

Model	Relative Error	5% IOD	10% IOD	20% IOD
CompASM	0.0769	46.33%	77.52%	94.49%
AR model	0.0353	79.93%	96.85%	99.80%

Table 3.4: Width and height error rate from CompASM and AR model. (© 2014, IEEE)

Component	Model	Width Error	Height Error
Right Eye	CompASM	0.0595	0.0308
	AR model	0.0265	0.0233
Left Eye	CompASM	0.0515	0.0300
	AR model	0.0252	0.0216
Mouth	CompASM	0.0979	0.1516
	AR model	0.0359	0.0399

Table 3.5: Relative error and detection rate from CompASM and AR model on each expression. (© 2014, IEEE)

Evaluation Metric	Neutral		Smile		Angry	
	CompASM	AR model	CompASM	AR model	CompASM	AR model
m17 landmarks	0.0541	0.0342	0.0735	0.0334	0.0550	0.0340
Right Eye Width	0.0528	0.0249	0.0616	0.0244	0.0459	0.0232
Left Eye Width	0.0401	0.0218	0.0483	0.0231	0.0405	0.0215
Mouth Width	0.0483	0.0318	0.1583	0.0459	0.0513	0.0321
Right Eye Height	0.0202	0.0228	0.0248	0.0244	0.0253	0.0226
Left Eye Height	0.0184	0.0221	0.0219	0.0210	0.0242	0.0234
Mouth Height	0.0444	0.0429	0.0530	0.0332	0.0459	0.0386
5% IOD	60.50%	80.83%	44.67%	82.09%	56.36%	81.83%
10% IOD	87.55%	96.69%	77.27%	97.30%	88.71%	97.48%
20% IOD	98.74%	99.84%	95.39%	99.84%	98.90%	99.84%



Figure 3.10: Samples of CompASM results on scream expression. The facial landmarking performance is not so accurate on scream expression.

Table 3.6: Relative error and detection rate of the AR model on various colour spaces.
(© 2014, IEEE)

	Relative Error	5% IOD	10% IOD	20% IOD
RGB	0.0391	76.01%	95.59%	99.66%
HSV	0.0401	75.02%	95.47%	99.53%
Grey	0.0393	76.08%	95.47%	99.57%
RGB-NII	0.0405	74.95%	94.91%	99.43%

3.3.4 The AR Model with Different Colour Spaces

In this section, we will use different color models to train and test the face images for landmark detection. This is motivated by the existing research in face recognition as different color models will have significant impacts. We did similarly on the proposed AR model. The relative error and detection rate are summarized in Table 3.6. The highest performance was achieved in both RGB and grey-scale colour spaces. However, the performance differences are not too significant and we cannot draw a clear conclusion on this issue as in the face recognition. Variation in colour spaces does not seem to strongly affect the result of facial landmarks detection. This can be explained by the fact that the edge information are still preserved well on the chosen colour spaces. Since the AR model utilizes HOG features which only rely on edge information, it still performs relatively similar. This is different from the scenarios of face recognition.

3.4 Summary

In this chapter, we proposed a new landmark detector, **the AR Model**, based on the work by Zhu and Ramanan. The proposed **AR model** is derived from a more sophisticated face structure by adding a high level of landmark density for possible better semantics descriptions. The AR model contains 130 landmarks trained on AR database which is almost twice as many as the Independent-1050 can provide. In the process of building this new model, we use AR database due to the availability of large amount of landmarks ground truth.

We employed experimental setup to measure the error rate of landmarks, detection rate, and geometric description accuracy in order to compare our proposed model with other two face models: the Independent-1050 and the CompASM. The results show that our proposed model outperforms both of them significantly. In fact, these results confirmed

our expectation for this new detector and we will use it in future study in the next few chapters. The last experiment investigates the effect of colour spaces on the AR model which shows no significant correlation between detection accuracy and chosen color space.

Despite the significant increase in performance, the AR model can detect the landmarks properly only on large faces. In practice, as the face image resolutions are not unique, we will investigate how to develop landmark detectors for different resolution of images, especially for faces in very low resolutions. This issue will be investigated and new face models will be proposed in chapter 4.

Chapter 4

Facial Landmarks Detection for Multi-Resolutions Images

In real-life applications, there is no guarantee that the size of the faces in an image are always in high resolution. We need to consider the scenarios where the face images are in lower resolutions which can be caused by a large distance between persons and cameras. Since the proposed AR model in last chapter is trained on high resolution faces, the learned features are only compatible with high resolution faces. In fact, if the minimum face size is below 240x240, it starts to fail in detecting the facial landmarks. Even the Share-146 face model (Section 2.1.4) developed by Zhu and Ramanan (2012a) is only capable to detect landmarks with resolution above 80x80. This is a significant limitation because it is still possible to conduct facial recognition on even smaller faces (Zhao *et al.*, 2003).

Motivated by this observation, we will propose the Multi-Resolutions (MR) models in this chapter to integrate with the proposed AR model (chapter 3) and expand the face resolution from high resolution down to 30x30 pixels for facial landmarks detection. Our investigation reveals that the initial amount of landmarks (130) from the AR model is too dense for small faces, which hinders the process of face models training. Therefore, we will develop an automatic adaptive landmarks scheme to achieve attentive selection of particular important facial landmarks. The facial landmarks are chosen accordingly based on the size of the face images in order to optimize computation time while providing sufficient landmarks as small faces contain less detail. Through this analysis, we aim to train 4 face models on various face resolutions: 210x210, 150x150, 90x90, and 30x30.

The performance evaluations are based on the *relative error rate* and *detection rate* with the same set of thresholds used in Section 3.3.1 on 11 fiducial landmarks. 196 Frontal face images from PUT database (Kasinski *et al.*, 2008) were selected as the testing dataset. Our performance evaluations involved three other facial landmarking approaches. The first one is the Share-146 model proposed by Zhu and Ramanan (2012a) since it was claimed to be able to detect the smallest faces among all their proposed face models. This model is compared as a baseline evaluation. The second one is the STASM developed by Milborrow and Nicolls (2014) which is an improvement of the Active Shape Model (ASM) approach

(Cootes *et al.*, 1995). They employed SIFT features (Lowe, 2004) in a simplified form integrated with Multivariate Adaptive Regression Splines (MARS) approach (Friedman, 1991) for efficient features matching. The last one is the Intraface approach proposed by Xiong and De la Torre (2013). They solved the optimization problem of Non-linear Least Squares (NLS) function by introducing their Supervised Descent Method (SDM) approach. The amount of landmarks detected by the STASM and Intraface are 77 and 49 respectively.

The structure of this chapter is organized as follows. Section 4.1 describes the details of framework setup for training the MR models including landmarks reduction and selection process via adaptive landmarks scheme. Section 4.2 outlines the detail of testing face images set and evaluation protocols used for performance comparison. The results will then be discussed and analyzed to assess the improvement of the MR models. Lastly, the summary is presented in Section 4.3.

4.1 The Multi-Resolutions (MR) models

As mentioned in section 2.1.4, Zhu and Ramanan published three of their proposed face models for research purposes (Zhu and Ramanan (2012b)). The first one is the Independent-1050 model used for comparison in chapter 3. The other two models are the Share-99 and Share-146 with parts-sharing trait for computational efficiency. According to their claim, the Share-146 was trained to detect landmarks on smaller faces compared to the other two models. the Share-146 can perform well on faces with size at least 80x80 pixels. However, since face recognition is feasible on even smaller faces (Zhao *et al.* (2003)), this fact motivates us to propose new face models of facial landmarks detection for smaller range of face image sizes.

Previously, we have proposed the AR model for better accuracy and geometric description. However, since it was trained on large face images from AR database (approximately 300x300 pixels) (Martínez and Benavente (1998)), it only performs well on high resolution face images. The required face size is at least 240x240 before it fails to perform landmark fitting properly. Based on this observation, we attempt to re-train different face models with similar tree structure and number of landmarks of AR model but on smaller size of face images. However, the data for small facial images and their landmarks in ground truth were not available and we have to obtain them by scaling down the same training data set used on the AR model by using the bicubic interpolation technique (Mat (2012)). We were able to complete the model training only on slightly smaller faces. As a matter

of fact, the reason behind this failure for the proposed AR model on high resolution is the high level of facial landmarks density imposed on small faces. Since the landmarks are too close to each other on small faces, most landmark features would lose their uniqueness and make it more difficult to distinguish between neighboring landmarks. Thus, we have to adjust the number of landmarks and structure of the face models to fit small size of face images. As the original AR model becomes a part of MR models only for high resolution faces with 130 facial landmarks, we refer it as the *MR-130 models* in the sequel.

4.1.1 Adaptive Number of Landmarks via Resolution Reduction

The first question for the number of landmarks for a given image should be solved first. Our extensive experiments show that the acceptable threshold for face size reduction is approximately 80% of the initial scale. Beyond this number will cause the face landmarks to be too dense for a successful training. Based on this observation, we propose an automatic and systematic framework to reduce less essential landmarks accordingly depending on the intentional training face size.

We need to consider **three essential aspects** in designing this landmarks reduction framework. **First**, since the detected landmarks will be used on face-related applications such as face recognition, there is a necessity on preserving important/fiducial facial landmarks in the reduction process. Our observation motivates us to include at least the nose tip and corners of both eyes and mouth. **Second**, we should retain the symmetrical proportion of the face tree structure by reducing the landmarks in a uniformly-distributed manner. For instance, it is undesirable to have 6 landmarks on one eye while the other eye contains 10 landmarks. **Third**, all the landmarks have to be rearranged once a landmark is removed. Otherwise, it will leave a large *gap* between neighboring landmarks (Figure 4.1).

In order to preserve the primary/fiducial landmark points, we select some "special" landmarks with high priority. These landmarks will not be removed from the face model structure with any face resolution. We refer them as the *Very Important Points (VIP)*. After much consideration, we decided to develop the face architecture based on the chosen VIP as shown in Figure 4.2. Inspired by the critical landmarks defined by Çeliktutan *et al.* (2013) and some additional landmarks, we selected 18 VIP i.e. 8 corners on both eyebrows and eyes, 1 nose tip surrounded by 3 landmarks on nose contour and 1 between the eyes, 2 mouth corners, and 3 along the face contour.

The role of VIP is not only limited to preserving crucial landmarks, but also to serve as

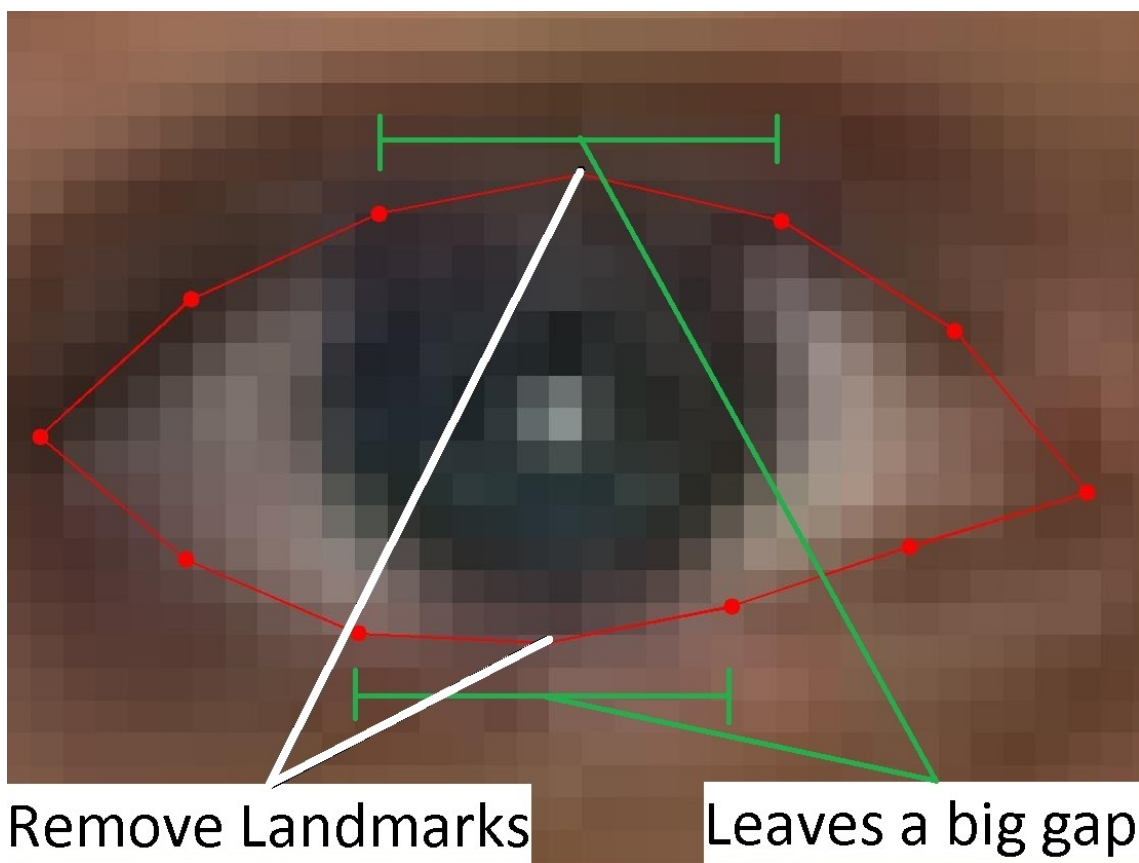


Figure 4.1: A large gap created every time a landmark is removed. (© 2016, AIMS)

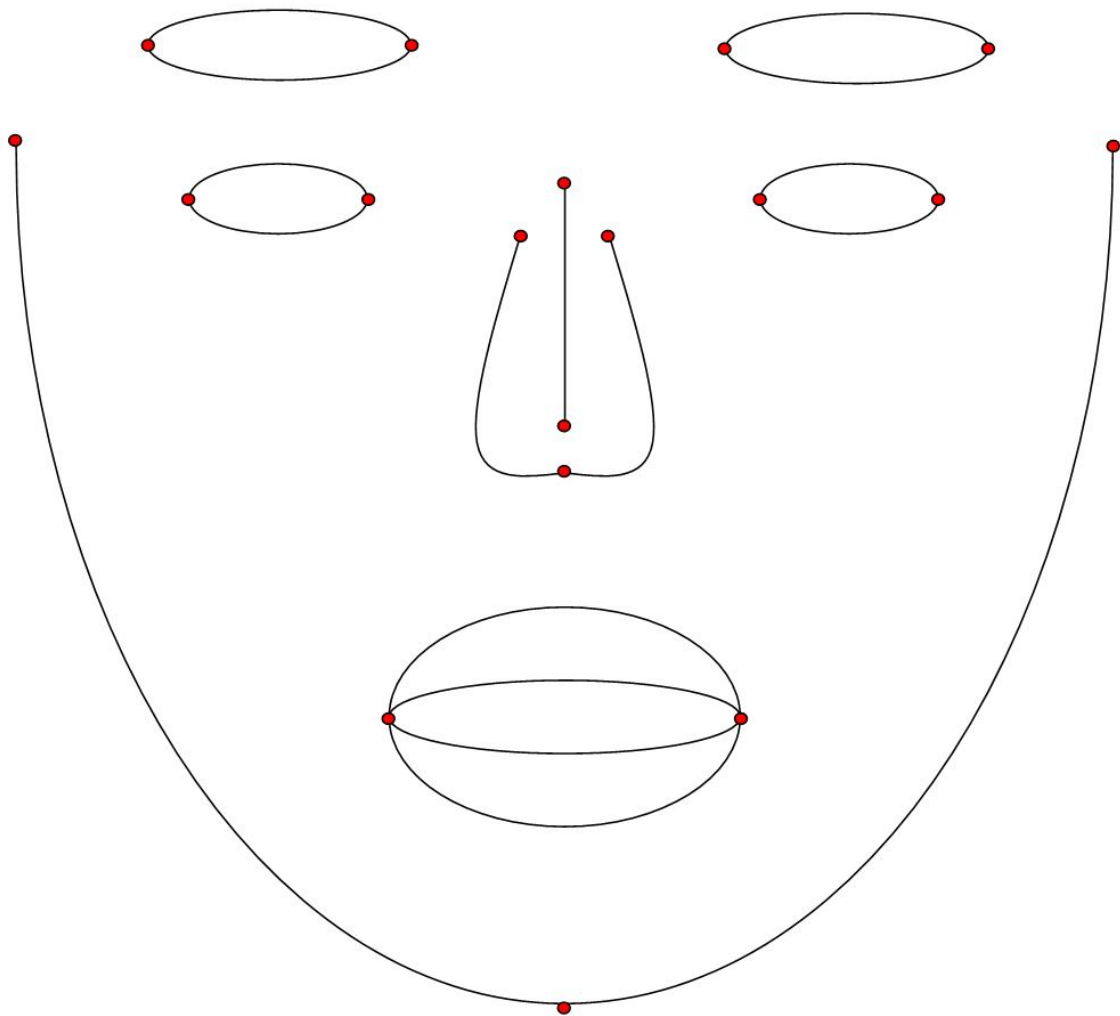


Figure 4.2: 18 chosen Very Important Points (VIP) to preserve on the proposed MR models. (© 2014, IEEE. 2016, AIMS)

a border to split the face structure into separate *segments*. The adjustment was done to ensure each segment is enclosed by two VIP. The amount of facial landmarks in-between will be reduced gradually while the face resolution for training gets smaller. The reason for this scheme is to have a balanced and symmetric reduction on the whole face structure. In total, we have 17 segments as shown in Figure 4.3.

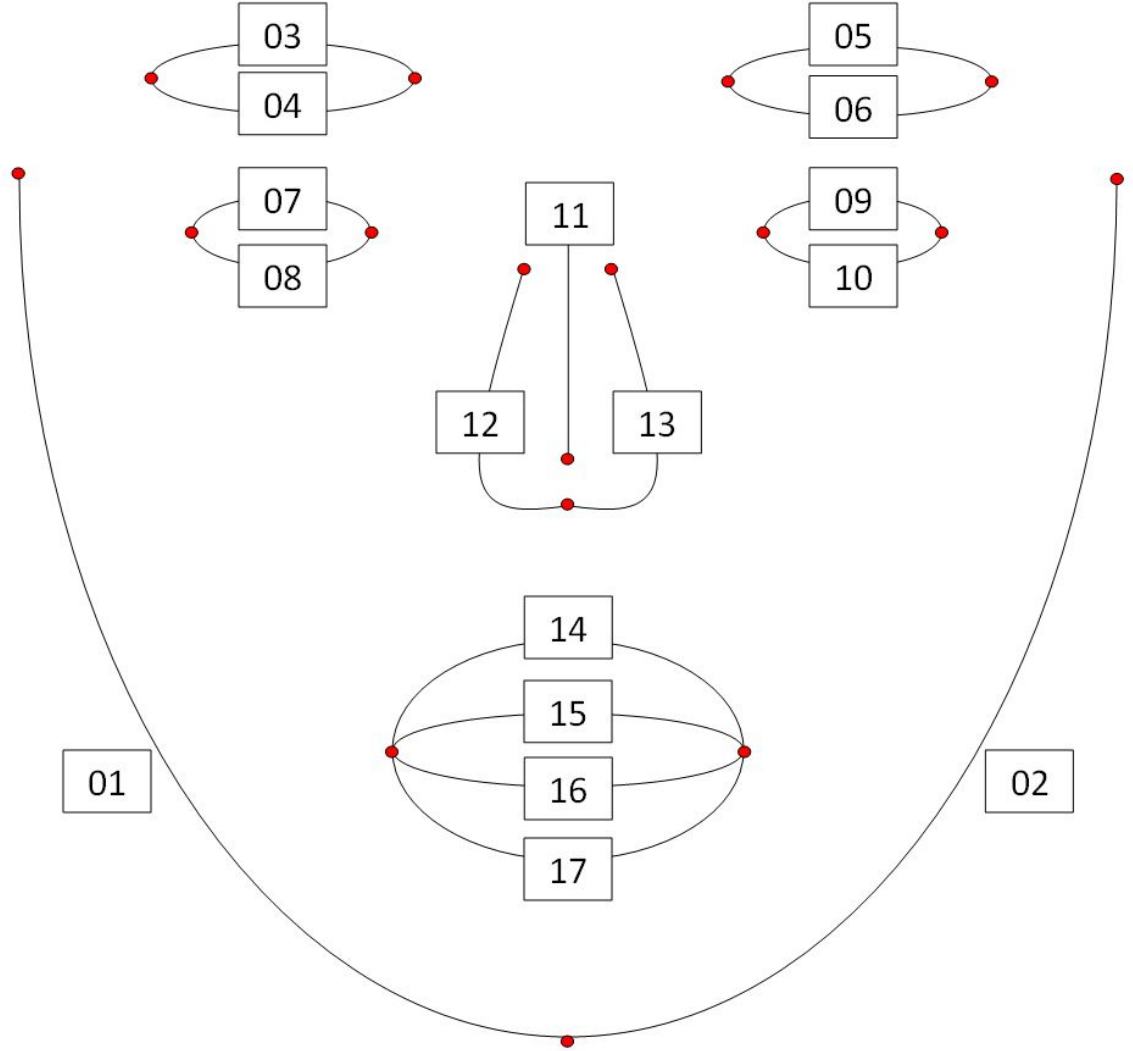


Figure 4.3: Face structure is divided into 17 segments by the VIP. (© 2014, IEEE. 2016, AIMS)

In the resolution reduction process, since reducing a landmark would leave a trace of *gap*, we need to adjust the new position of the remaining landmarks to maintain a consistent distance along the initial line of landmarks. To visualize this scenario, refer to an example in Figure 4.4. Two segments (upper and lower eye lid) between two VIP (eye corners) are shown along with the initial eye landmarks represented with red dots. If one landmark is removed on each segment, the new set of landmarks are rearranged as shown by the green

dots. As can be seen in this example, the revised landmarks stay on the initial trajectory of the eye contours to preserve the geometric detail as much as possible. This process is repeated until the distance between neighboring landmarks is greater than initial 80% scale distance. Finally, the whole procedure to reduce landmarks in a single segment is summarized in algorithm 4.1.

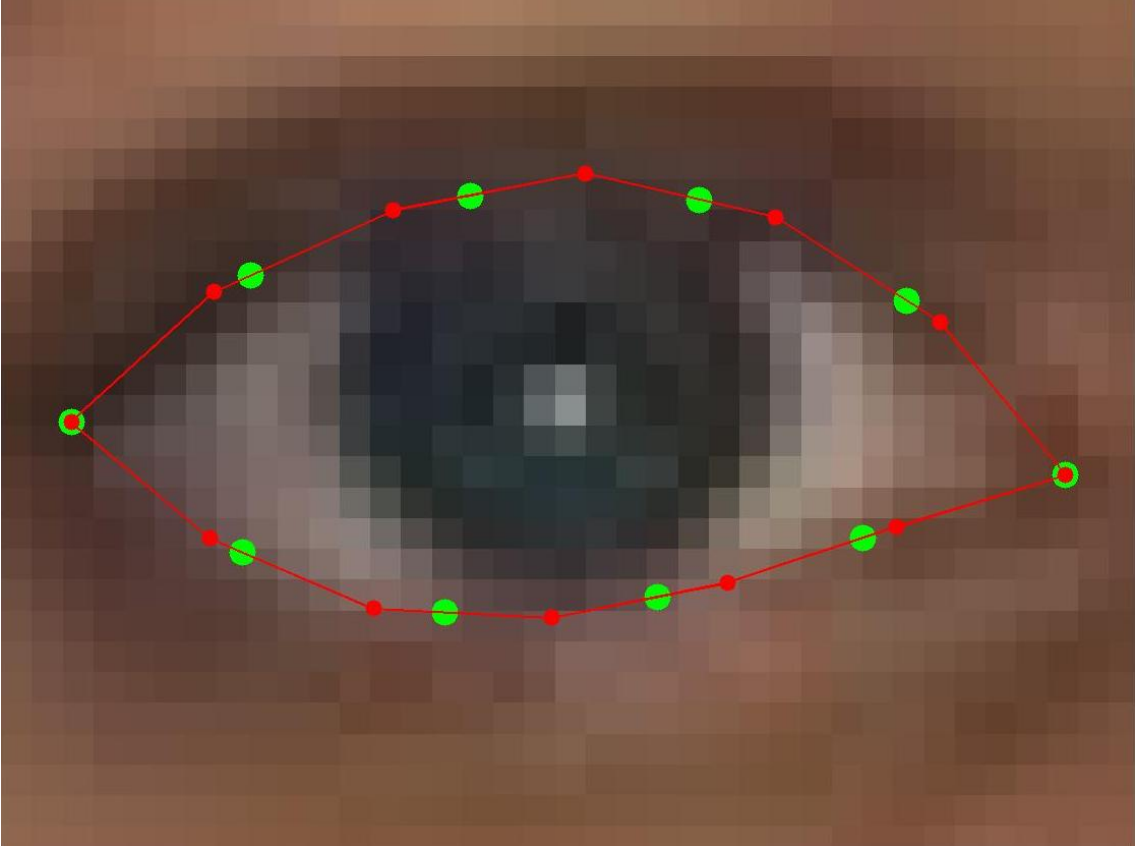


Figure 4.4: Rearrangement of the landmarks when performing landmarks reduction. Red dots represent the initial landmarks while the green dots represent the revised landmarks. (© 2016, AIMS)

4.1.2 Training for the MR Models

With the proposed landmarks reduction schema to produce less landmarks density in the face images training set, we can train the MR models on any resolution of choice after proper reduction. The face images for training are from the AR database by Martínez and Benavente (1998) with the corresponding landmarks ground truth (Ding and Martínez (2010)). We also use 4 facial expressions (neutral, smile, angry and scream) from 112 subjects each, the same set used for model training in section 3.2. We decided to

Algorithm 4.1 Landmark Reduction on Each Segment between 2 Very Important Points (VIP). (© 2016, AIMS)

1: **procedure** REDUCE_LANDMARKS(coord, percentage)

coord is a set of coordinates (x,y) of landmarks.

the first and last landmarks are the VIPs.

percentage is the scale of the landmarks based on the corresponding face image.

the assumption is that the percentage is less than 80% image scale.

final_coord \leftarrow []

```

2:   row  $\leftarrow$  size(coord, 1)                                ▷ row is the amount of landmarks
3:   landmarks  $\leftarrow$  row - 1
4:   total_percentage  $\leftarrow$  percentage * 100.0 * (landmarks)
5:   dist_per_pair  $\leftarrow$  total_percentage / landmarks
6:   while dist_per_pair < 80 and landmarks  $\neq$  1 do ▷ reduce the landmarks one by
   one
7:     landmarks  $\leftarrow$  landmarks - 1
8:     dist_per_pair  $\leftarrow$  total_percentage / landmarks
9:   end while
10:  if landmarks = 1 then
11:    final_coord  $\leftarrow$  [coord(1) coord(row)]                ▷ only 2 VIPs remain
12:  else
13:    final_coord  $\leftarrow$  [coord(1)]
14:    for i = 1 to landmarks do
15:      position  $\leftarrow$  (i / landmark) * (row - 1)
16:      int  $\leftarrow$  floor(position)
17:      frac  $\leftarrow$  position - int
18:      if int + 1 < row then
19:        new_landmark  $\leftarrow$  coord(int + 1) + (coord(int + 2) - coord(int + 1)) * frac
20:      else
21:        new_landmark  $\leftarrow$  coord(int + 1)
22:      end if
23:      final_coord  $\leftarrow$  [final_coord new_landmark]
24:    end for
25:  end if
26:  return final_coord
27: end procedure

```

train 4 MR models on 4 scale levels: 70%, 50%, 30%, and 10% (image resize via bicubic interpolation approach (Mat, 2012)). These models cover various resolutions down to size 30x30. Furthermore, four facial expressions (neutral, smile, angry, scream) are involved on each scale. The process of landmarks reduction for each scale level can be observed in Figure 4.5. It shows how the landmarks is gradually being reduced as the face size gets smaller. For a better visualization purpose, the face images in the figure were not scaled down in order to emphasize the landmarks reduction process. However, the real model training used the scaled-down face images.

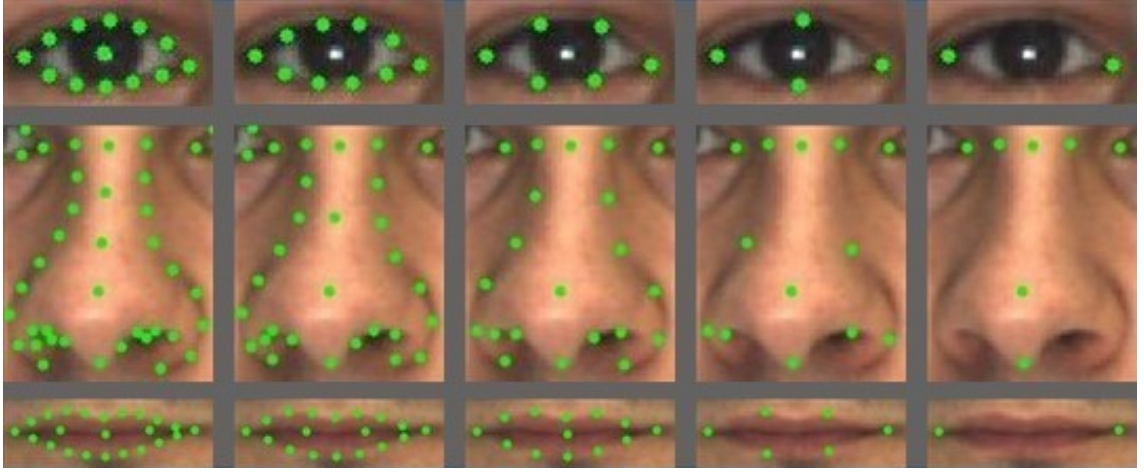


Figure 4.5: The landmarks reduction process on a face. We emphasize on the facial components eye, nose, and mouth in this figure. The order of the scale level is as follows: 100% (ground truth), 70%, 50%, 30%, 10%. The face images were not scaled down here for easier observation. (© 2014, IEEE)

As mentioned previously, the previous AR model is now referred as the MR-130 model because it has 130 landmarks. Following the same naming principle, the other MR models are named the MR-103, MR-70, MR-36, and MR-14 for 70%, 50%, 30%, and 10% scale levels respectively. The information on the **MR models** is **summarized** in Table 4.1. The complete set of MR models are shown in Figure 4.6.

Table 4.1: The summary of the MR models. All of them are trained on four facial expressions (neutral, smile, angry and scream) from 112 subjects from AR database. (© 2014, IEEE. 2016, AIMS)

MR Models	Target Face Sizes	Landmarks Amount	Training Face Sizes (Approximate)
MR-130	Above 255	130	300x300
MR-103	180 - 255	103	210x210
MR-70	120 - 180	70	150x150
MR-36	60 - 120	36	90x90
MR-14	Below 60	14	30x30

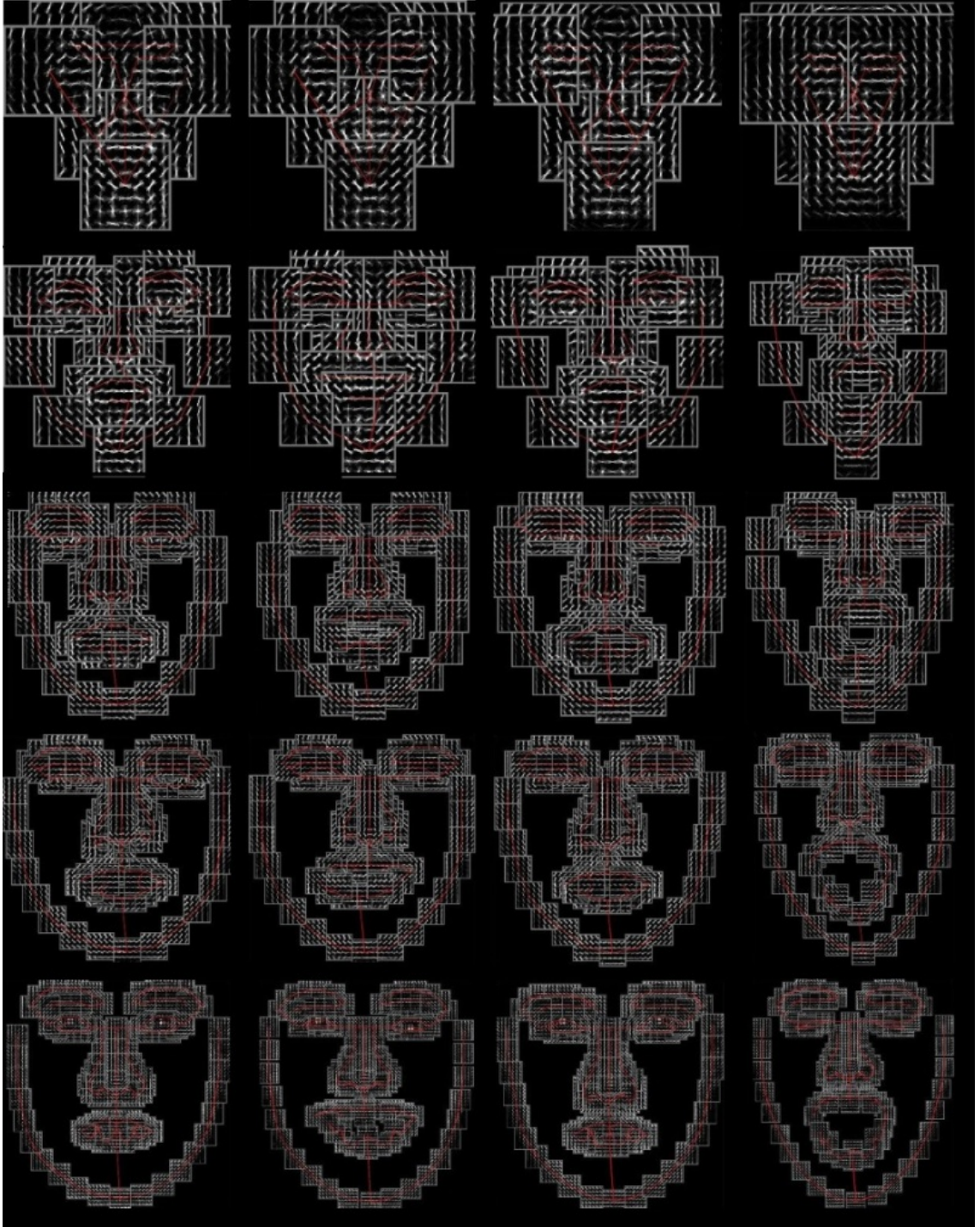


Figure 4.6: The complete set of MR models. Starting from the first row are the MR-14, MR-36, MR-70, MR-103, and MR-130. Various facial expressions are shown in the order of neutral, smile, angry, and scream. (© 2016, AIMS)

We should make a small exception of the VIP rule on the MR-14 model. As there are 18 VIP, the MR-14 should at least contain 18 important landmarks. However, we decided to keep only one landmark on the nose tip and ignore the other 4 landmarks around nose region since the features are too subtle on a very small faces according to our observation. Refer to Figure 4.7 for the visualization.

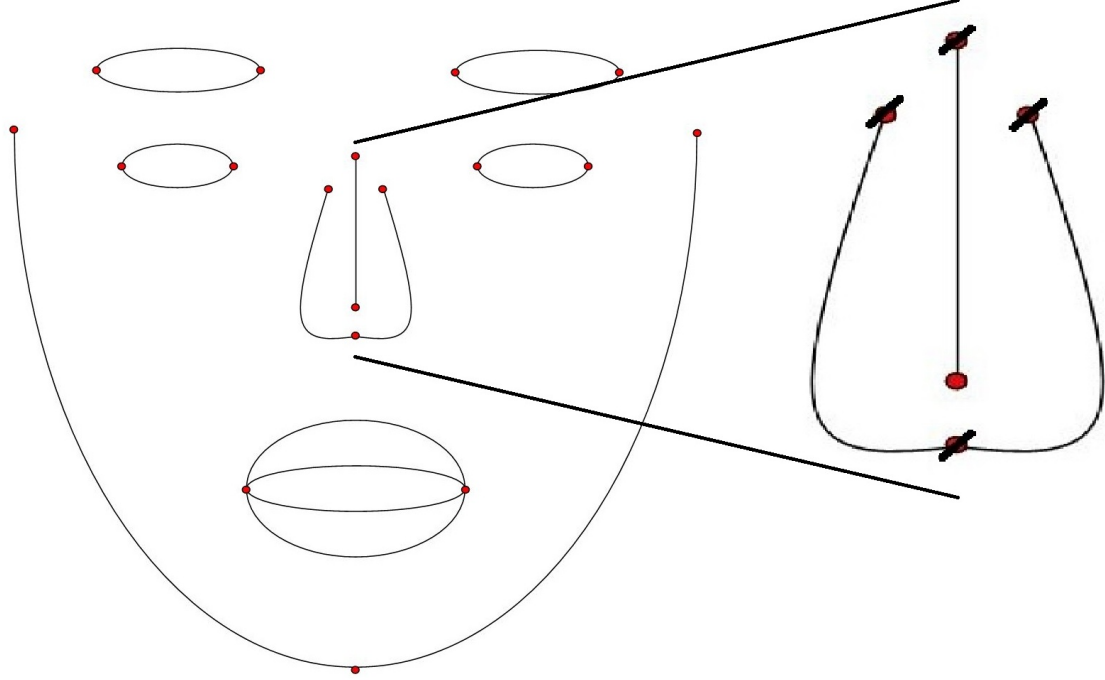


Figure 4.7: 4 VIP are exempted from the MR-14 model. Our observation shows that the features are too subtle to be included. (© 2016, AIMS)

4.2 Experiments

For the different AR models developed in last section, we first need to measure the performance improvement of our proposed MR models by comparing with the Share-146 model by Zhu and Ramanan (2012a). The reason for choosing the Share-146 is because this model can reach the lowest resolution (down to 80x80) compared to the other proposed models. We then compared it with two other robust facial landmarking approaches: the STASM model by Milborrow and Nicolls (2014) and Intraface model by Xiong and De la Torre (2013). Next, we will address the experiments in detail.

4.2.1 Testing Dataset

As all the models are trained by the AR dataset, we use the PUT dataset for our testing in this chapter for fairness to different detectors. Frontal faces from PUT database (Kasinski *et al.* (2008)) (section 2.5.1.3) were used for evaluating performance of the proposed MR models as the ground truth for some important landmarks are given in this dataset. All the images are available in high resolution 2048x1536 with face sizes approximately 750x750 on controlled illumination. Since the faces are provided in a sequence of rotating head (various poses), we had to manually choose frontal faces out of the first two face pose subsets. In total, 196 frontal face images from 98 participants were selected as the testing set.

Besides the original face size (750x750), we also scaled it down to seven various sizes to test each MR model (Figure 4.8). We adjusted the scale level accordingly to gain face sizes on approximately 600x600, 450x450, 300x300, 210x210, 150x150, 90x90, and 30x30. Next, we will present the performance evaluations with different detectors.

4.2.2 The Evaluation Protocols

Some evaluation protocols used in section 3.3.1 are applied in this experiment. To be more specific, we measured the *relative error rate* and *detection rate* with the same set of thresholds. However, we only compare 11 landmarks due to fewer common landmarks between the proposed MR models and other models (Figure 4.9). Furthermore, measuring accuracy of geometric descriptions is not included as there are not sufficient details of facial components on low resolution faces.

4.2.3 The MR Models VS the Share-146 Model

We first focused on the results on the comparison with the Share-146 model by Zhu and Ramanan (2012b) as the baseline evaluation. The results are summarized in Table 4.2 and 4.3. For large faces (300x300 or above), our proposed MR models produce approximately **40% less error rate** and **30% more detection rate** for the lowest threshold (5%). The MR models also still outperform the Share-146 on the other thresholds. For the cases of small faces (with resolution of 210x210 or lower), even though the performance gap is less, the improvement is still apparent. The impact is more obvious especially on the case of the smallest face (30x30) where the Share-146 cannot even detect the presence of the

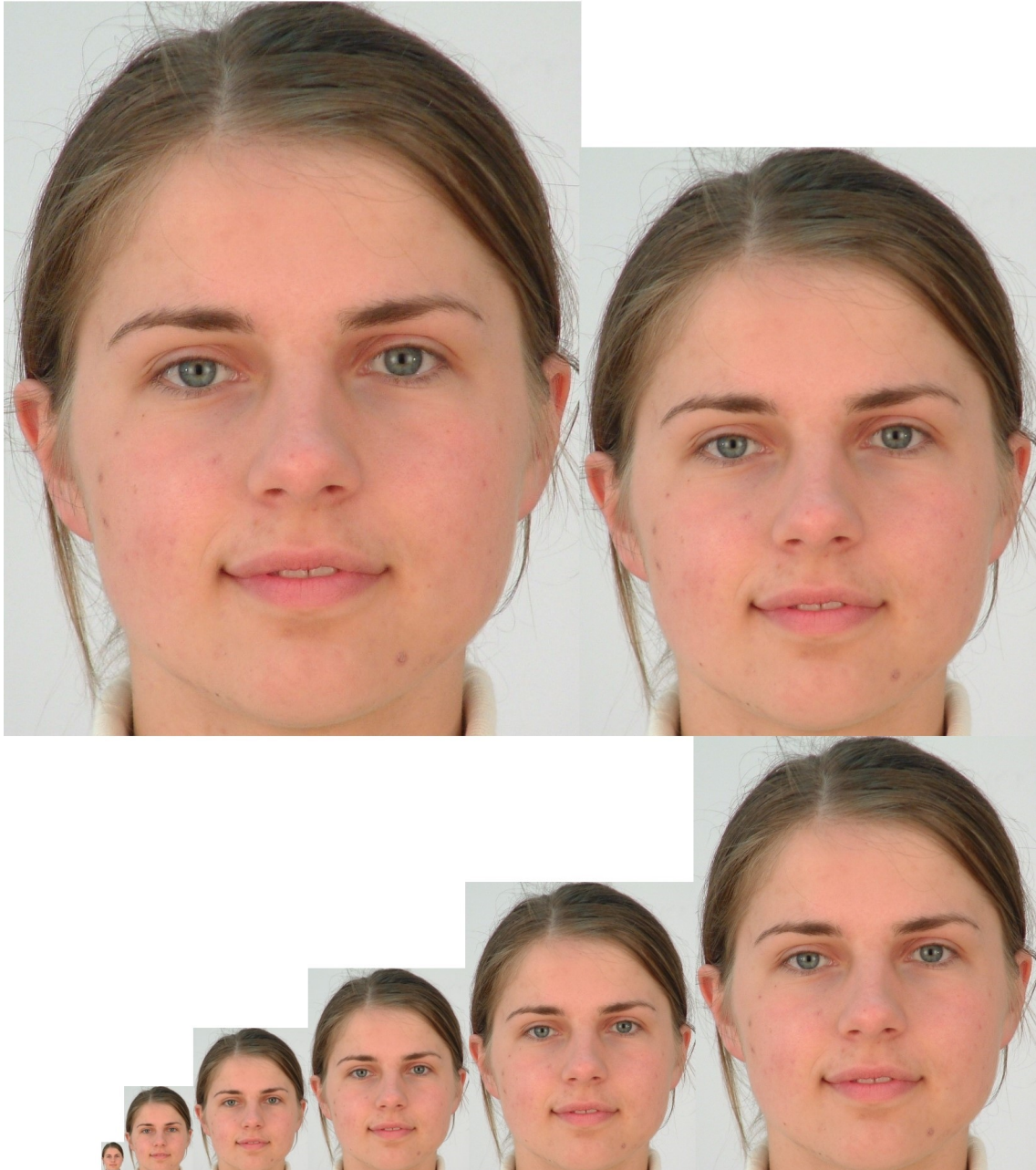


Figure 4.8: Samples of face image in various resolutions. In clockwise direction, the sizes shown here are 750x750, 600x600, 450x450, 300x300, 210x210, 150x150, 90x90, and 30x30. It is clearly seen that the information difference between large and small faces are imminent. (© 2014, IEEE)

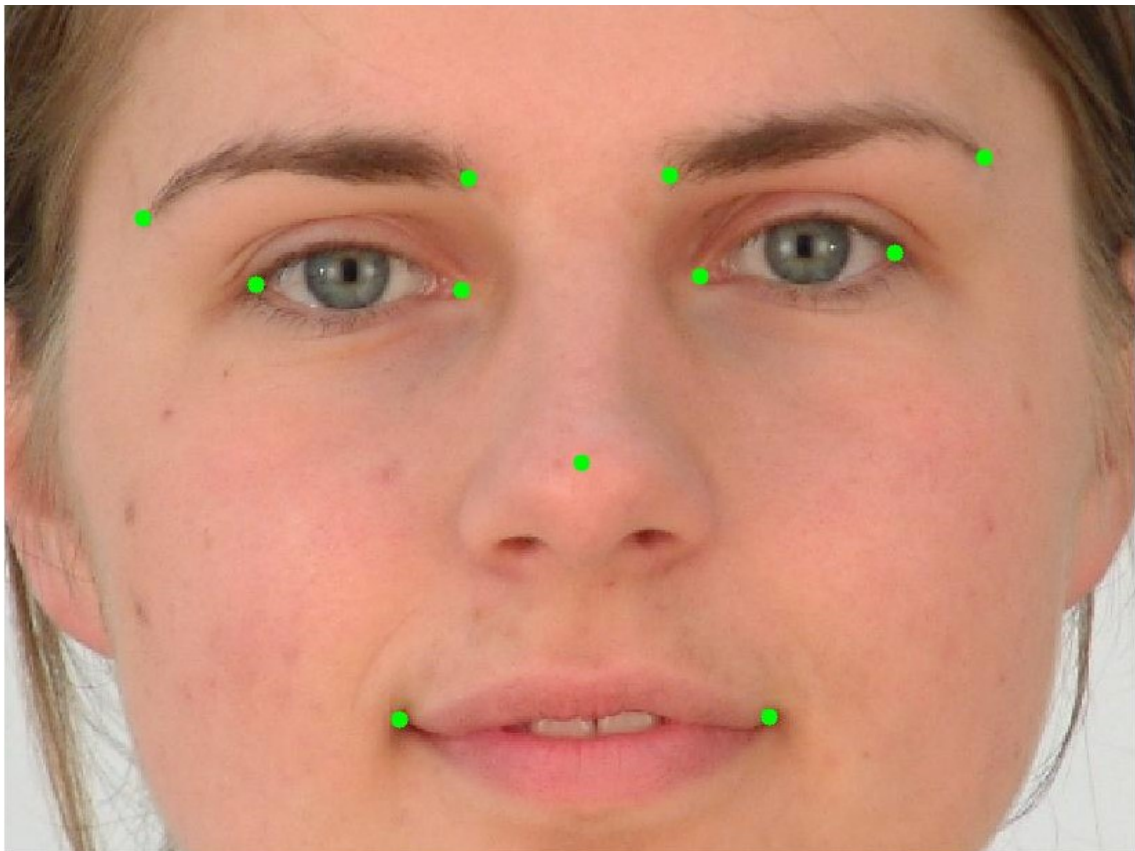


Figure 4.9: Eleven facial landmarks for performance evaluation. (© 2014, IEEE. 2016, AIMS)

faces, which is the main disadvantage of Zhu and Ramanan’s face models.

Table 4.2: 11 Facial Landmarks Relative Error from the SHARE-146 model and MR Models. (© 2014, IEEE. 2016, AIMS)

Face Sizes	SHARE-146	MR model
750x750	0.0911	0.0543
600x600	0.0914	0.0542
450x450	0.0912	0.0537
300x300	0.0922	0.0548
210x210	0.0902	0.0596
150x150	0.0930	0.0669
90x90	0.0920	0.0832
30x30	not detected	0.1225

Table 4.3: Detection Rate (%) from the SHARE-146 Model and MR Models. (© 2014, IEEE. 2016, AIMS)

Face Sizes	5% IOD		10% IOD		20% IOD	
	S-146	MR	S-146	MR	S-146	MR
750x750	22.26	53.06	61.73	90.44	97.26	99.63
600x600	21.10	52.32	61.60	90.35	96.94	99.68
450x450	22.96	54.36	61.32	90.63	96.94	99.81
300x300	21.10	52.08	60.62	89.70	96.75	99.77
210x210	22.87	45.55	61.87	87.48	97.31	99.86
150x150	20.83	38.68	59.69	82.24	96.85	99.35
90x90	20.22	25.42	61.64	68.92	97.36	98.61
30x30	-	12.76	-	41.05	-	86.97

4.2.4 The MR Models VS Other State-of-the-art Approaches

In this section, we will compare the performance of the MR models and two other facial landmarking approaches: the STASM (Milborrow and Nicolls, 2014) and the Intraface (Xiong and De la Torre, 2013) with the Share-146 as the baseline performance. For an easier comparison and better visualization, we summarized the results of the error rate and detection rate as a line graph in Figure 4.10, 4.12, 4.13, and 4.14.

First, we observed the error rate in Figure 4.10. The MR models produces a slightly higher error rate on low resolution faces, but perform really well on high resolution faces on a par with the Intraface. The interesting result in this graph is the fact that the STASM provides facial landmarks with the least error on small faces but extremely high error rate on large faces. This seems unlikely to happen since large faces contain better information, thus lead to more accurate facial landmarking as shown by both the MR models and the

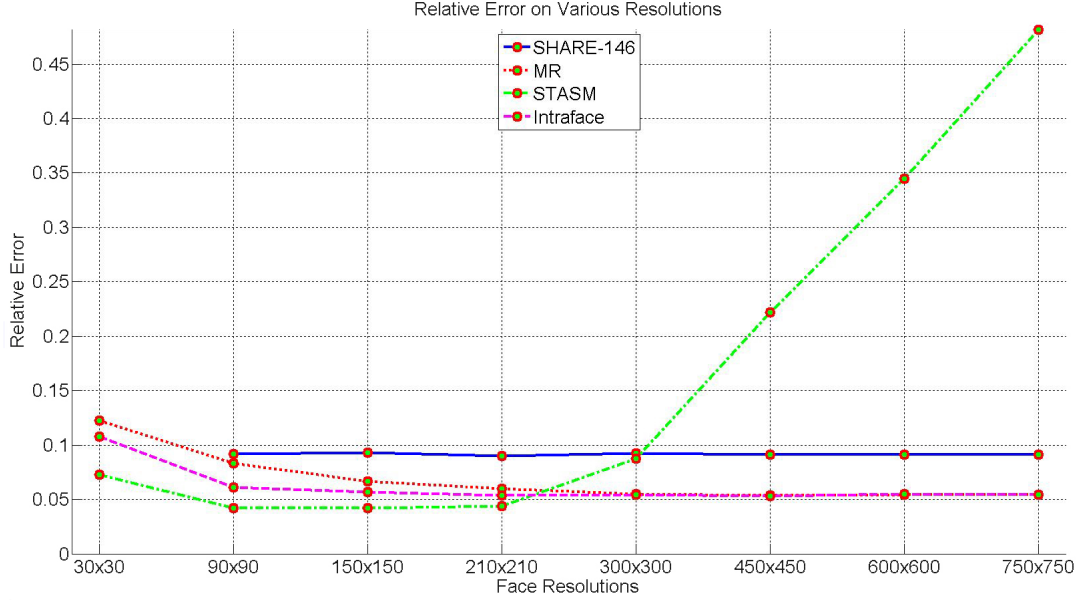


Figure 4.10: Relative error on PUT database.

Intraface. Since this is a peculiar case, we conducted a thorough examination and analysis on the STASM. Our investigation revealed that the reason behind the large error rate is its poor reliability on the accuracy of face detector.

Both the STASM and Intraface employed a well-known face detector approach by Viola and Jones (2004) (section 2.2). For computation efficiency purpose, the STASM and Intraface only attempt to detect facial landmarks on the region of face candidates detected by the Viola Jones face detector. However, unlike the Intraface, the STASM does not handle false detection well. The STASM imposes the facial landmarks even on the non-faces region. As a consequence, it has tendency to produce outliers on the performance as the detected facial landmarks might be located far from the facial components. Some examples can be seen in Figure 4.11. Its performance could easily get worse on large images since there is a higher chance to have more false face detection rate. Furthermore, this scenario still happens even with the fact that all testing face images were taken in a controlled environment.

The MR models and Share-146 can be applied without employing other face detectors since they explore the whole image and select the best face candidate based on the highest model score matching. However, for the sake of comparison, we also employed the Viola Jones face detector first to see how well it can handle false detection. Our experiment shows that there is no change on the performance (except for the faster computation speed) which implies that both the MR models and Share-146 can distinguish faces and

non-face regions well. They successfully ignored all the non-face regions due to not passing the landmarks fitting score.

The next thing we observe is the detection rate on various thresholds. We start from the smallest threshold 5% IOD in figure 4.12. It shows that the MR models detect landmarks slightly less accurate compared to the other two approaches. However, the other two cases of thresholds shown in figure 4.13 and 4.14 reveal the MR models' comparable performance and even better on high resolution faces. This is a strong indication that our proposed MR models are still able to detect approximate locations of facial landmarks well in a general situation, instead of imposing the landmarks on the ideal location but risking a total misalignment.

The Intraface shows a consistent and stable performance with slight increasing performance gradually along with the size of the faces. On the other hand, the STASM once again displays a gradual declining detection rate as it reaches large size of images, the same phenomena happened on the relative error rate. As expected, the false face detection rate from the Viola Jones face detector affects the detection rate significantly. Even though the STASM gives the best result on low resolution faces, the performance is not stable and are easily influenced by false face detection from the the Viola Jones for high resolution images.

In order to further demonstrate that the MR models are more robust against misalignment, we conducted a thorough examination to discover some of the examples on 30x30 faces where the Intraface and STASM encountered a significant issue as shown in figure 4.15. By allowing a landmark on the chin to stretch over, it assists fitting other facial landmarks of the MR models in the presence of beard or hair covering the eyebrows. Even though the location of the detected landmarks might not be perfect, it is compensated by fitting them on locations which are not too far off the mark to avoid full misalignment. We believe this happens as our proposed adaptive landmarks scheme can only preserve some important landmarks of facial components via reduction. On the other hand, the Intraface detects the beard as the part of the mouth, thus totally shifting mouth landmarks down to chin region. Furthermore, a slight occlusion around eyebrows creates a small-scale distortion on the upper landmarks by the Intraface and a total landmarks disorientation by the STASM.



Figure 4.11: STASM is susceptible to detecting facial landmarks incorrectly if it is employed on the non-face regions. Face detection with very high accuracy is required in this case. (© 2016, AIMS)

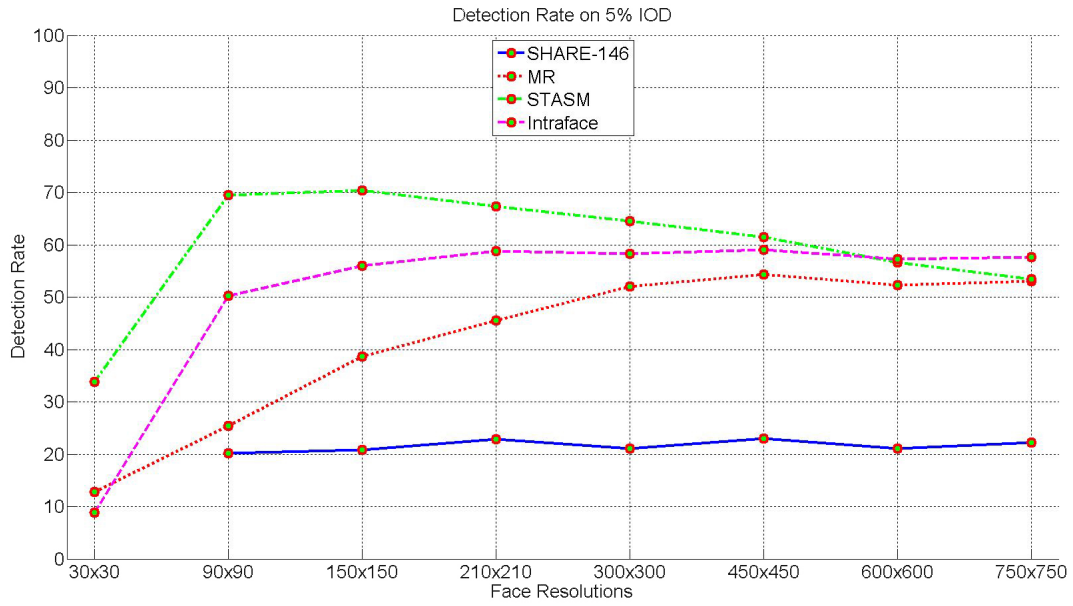


Figure 4.12: Detection rate on 5% IOD threshold.

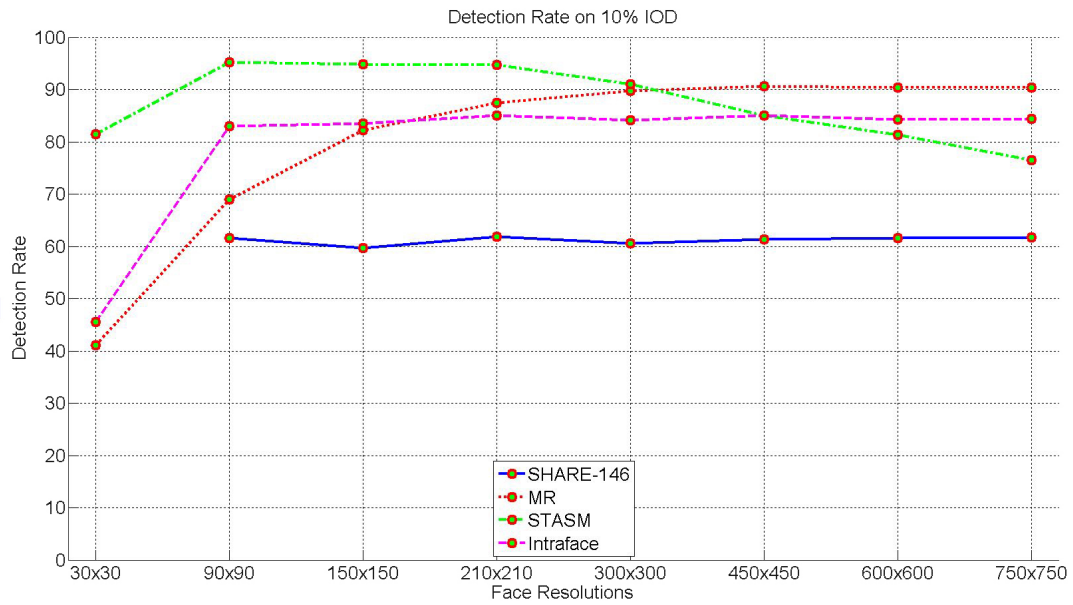


Figure 4.13: Detection rate on 10% IOD threshold.

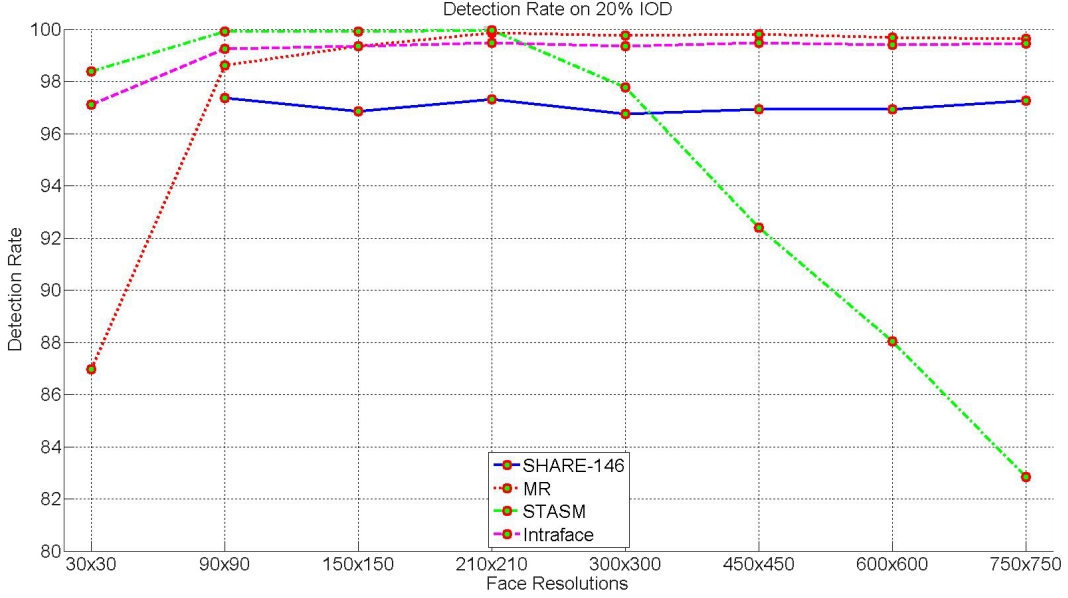


Figure 4.14: Detection rate on 20% IOD threshold.

4.3 Summary

In this chapter, we proposed the Multi Resolutions (MR) models. The aim is to be able to detect facial landmarks on small faces down to 30x30 since the previously developed AR model was trained only based on large faces (approximately 300x300) and will fail for small face images. As the original facial landmarks ground truth are too dense to fit on small faces, we proposed an automatic adaptive landmark scheme to select the important facial landmarks on various scales of face sizes. This allows us to train various face models for various face resolutions while maintaining adequate amount of facial landmarks. We chose to train the MR models on four sizes: 210x210, 150x150, 90x90, and 30x30.

The experiments were tested on 196 frontal face images from PUT database. The performances were evaluated based on the error rate and detection rate of 11 important facial landmarks. The first comparison is done on the Share-146 model as the baseline evaluation. Our MR models outperform the Share-146 by significant margin. In addition, the MR models are able to detect facial landmarks on the smallest face images of 30x30 on which the Share-146 is incapable. Further experiments were conducted on two state-of-the-art approaches, the STASM and Intraface. The results show that the MR model is slightly less accurate on small faces, but comparable on large faces. Additional experiments also show that the MR models are more robust against landmarks misalignment on the presence of beard and hair. Even though the STASM gives the best accuracy on small

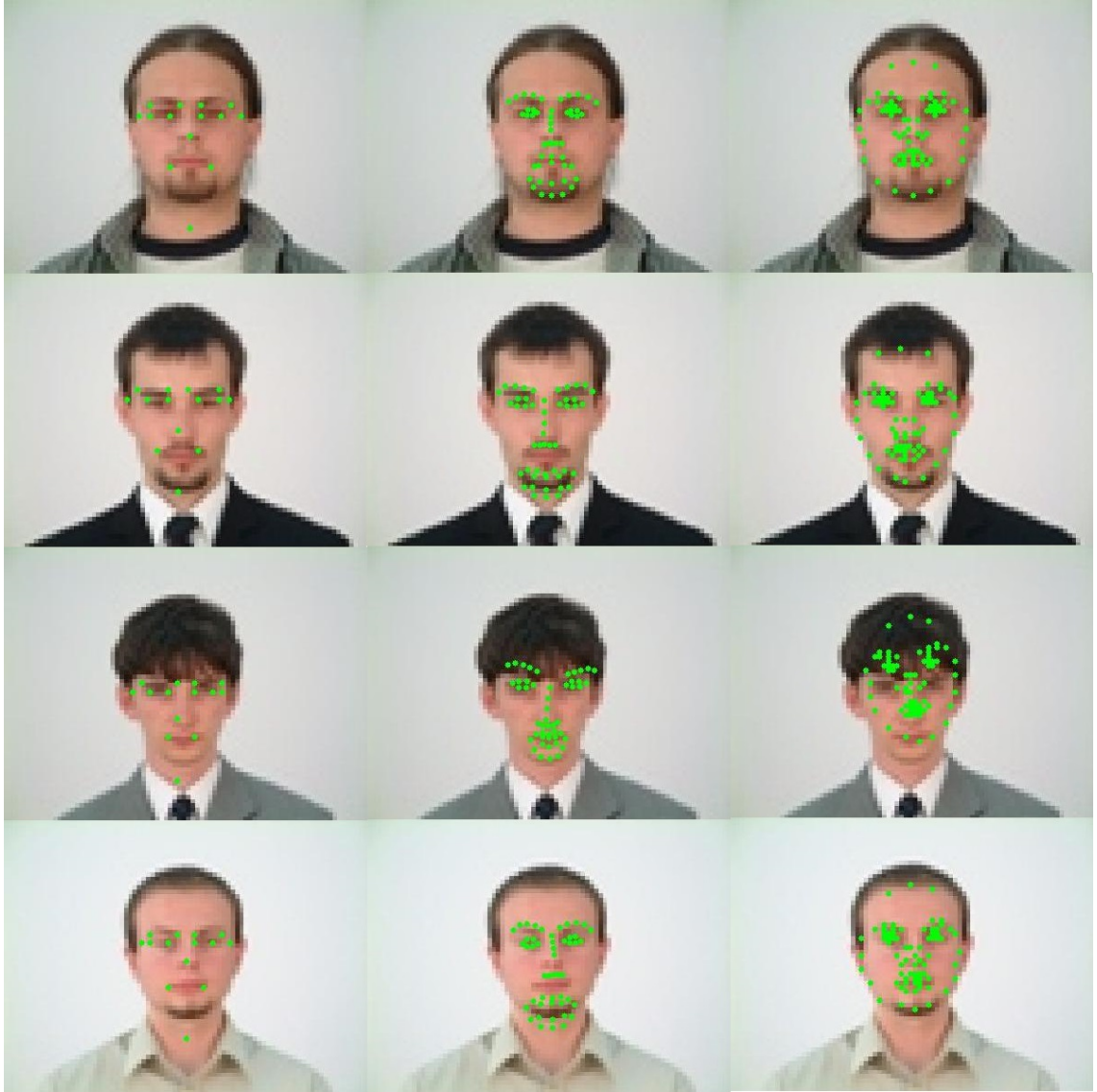


Figure 4.15: Examples of facial landmarks misalignments which occur on Intraface (**second column**) and STASM (**third column**) with 30x30 faces. Despite slightly less accurate, MR models have a major advantage of robustness against misalignment on facial components especially with the presence of beard and slight occlusion of eyebrows. (© 2016, AIMS)

faces, it is very sensitive on false face detection rate which leads to rapid fall on accuracy for large face images.

Despite the capability of the MR models to extract facial landmarks, it has only been tested on face images taken in controlled environment. The quality of these images are clear without significant noises/background which reduces the chance of false face detection rate. Furthermore, it is also known that each image only contains one face which makes the facial landmarking approach a bit easier since it just needs to select the best landmarks fitting (score). This ideal scenario will not happen on images taken on uncontrolled environment where the background might hinder face detection rate and there might multiple faces present on various resolutions. Therefore, we will investigate this issue by proposing a better face detection approach to act as pre-processing phase prior to facial landmarking in chapter 5.

Chapter 5

Fast and Effective Face Detector

The face detector approach developed by Viola and Jones (2004) is well-known to be robust and efficient due to its effective features and practical framework design (Section 2.2). This approach has been widely employed by some face-related applications including the STASM (Milborrow and Nicolls, 2014) and Intraface (Xiong and De la Torre, 2013) in the previous chapter. However, even with its real-time and accurate detection of "promising" face regions, the Viola Jones detector is still prone to high false positives as observed in last chapter in some controlled situation with different resolutions and this will be even worse on the uncontrolled environment.

As discussed in chapter 4, the MR models are capable of discovering the location of faces by performing full-scale scanning of the whole image, the similar approach is used in the Share-146 face models by Zhu and Ramanan (2012a). However, this method requires high computational time and is at higher risk of detecting false positives especially on cluttered background regions. If we assume that the approximate location and size of face regions are known beforehand, we can reduce such false positive rates significantly as shown in this chapter. In fact, we can identify the approximate location in low resolution by using the MR models and this low resolution detector also will reduce redundant computation. Therefore, there is a need to employ a reliable and efficient face detector prior to facial landmarking phase. This motivates us to propose an alternative way to utilize the tree-structured face models for filtering false face detection. We refer this face detection model as the Tree-structured Filter Model (TFM).

The experiments were tested on face images taken in uncontrolled environment. We first evaluate the performance of TFM combined with the Viola Jones detector based on the face detection accuracy. We compare it with the Viola Jones face detector and Share-146. This evaluation was tested on Face Detection Data set and Benchmark (FDDB) database (Jain and Learned-Miller, 2010) which provides sophisticated ground truth information of face locations along with the source code to produce the Receiver Operating Characteristic (ROC) curve. Hereafter, we will integrate the Viola Jones & TFM with the previously proposed MR models as a complete facial landmarking framework. We conducted another experiment on Annotated Facial Landmarks in the Wild (AFLW) (Koestinger *et al.*, 2011a)

database. Due to our research scope being focused on frontal faces, we manually choose considerably large amount of images containing only frontal/near-frontal face(s) from both databases. Finally, an additional experiment was conducted to assess the impact of image size on the growth of computational time for our proposed integrated system compared with the Share-146 face models.

The structure of this chapter is as follows. Section 5.1 describes the methodology of training TFM along with the corresponding performance evaluation. We then combine the proposed TFM with Viola Jones detector and the proposed MR models as an integrated facial landmarking framework in Section 5.2. The summary is addressed in Section 5.3.

5.1 The Tree-structured Filter Model (TFM)

The idea of combining the TFM with the Viola Jones face detector was inspired by two observations. First, the Viola Jones face detector detects face with *high* true positive rate in *real-time*. However, it comes with a lot of false positive in some situations. Second, the Tree-structured face models such as the Share-146 or Independent-1050 by Zhu and Ramanan (2012a) can distinguish false face detection better, but with the cost of significantly high computational time. Therefore, we would try to combine the advantages of both approaches to compensate the shortcomings of each other. We first apply the Viola Jones face detector for real-time detection and apply the proposed TFM to discard false positives. For a better visualization, we will provide an example for application of the Viola Jones & TFM combined with the MR models in Figure 5.1.

Accurate landmarks detection is not the main purpose of TFM. Instead, it was developed to detect the *presence* of faces by attempting facial landmarks fitting. Furthermore, we would like to design TFM to be lightweight by utilizing low resolution training faces and restricted amount of landmarks with an aim to capture the *intuitive descriptions* of frontal human faces efficiently. All face candidates from the Viola Jones detector are scaled down to 40x40 prior to passing it to the TFM to avoid high processing overhead. If the face candidates pass the filtering selection, then it is ready for facial landmarking phase (in original size, not 40x40 anymore). The pseudo code for our proposed scheme is shown in algorithm 5.1. After extensive testing, we choose 3 sub-windows as the merging threshold for the Viola Jones detector to achieves high rate of initial true positives and -1.10 as the landmarks matching score threshold for the TFM to remove false detections while preserving most of the correct detections.

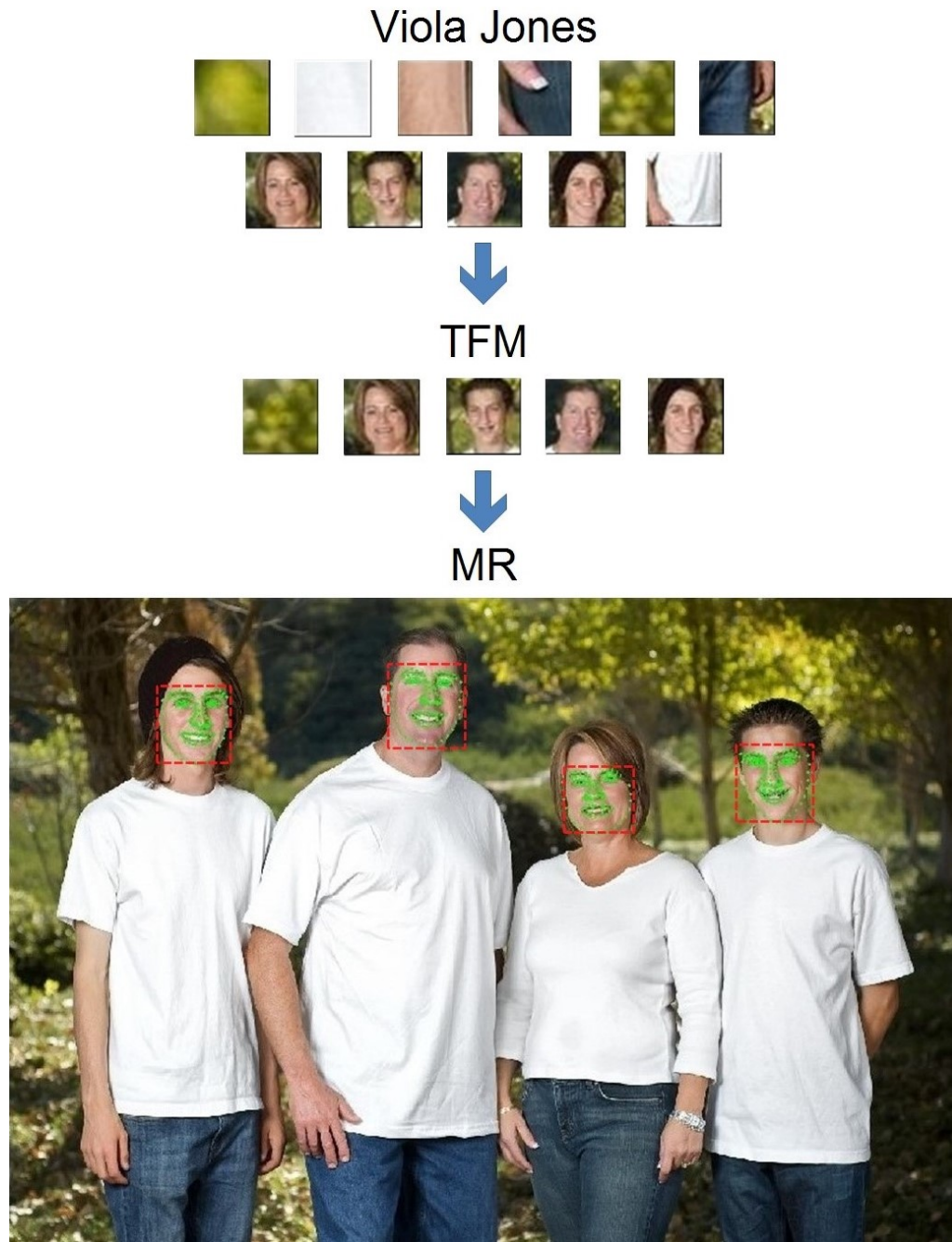


Figure 5.1: This is the illustration on how the Viola Jones (VJ) face detector performs together with Tree-structured Filter Model (TFM) concluded with facial landmarking by MR models. In this particular example, VJ successfully detect all 4 faces, but with the expense of 7 false positives. TFM then rapidly examines all the face candidates, successfully removing 6 false detections while maintaining all true detections. The last false positive is then disregarded by MR models. Since TFM has removed most of false detection quickly, it reduces the workload of MR models. (© 2014, IEEE)

Algorithm 5.1 Face Detection with Viola Jones (VJ) face detector and the proposed Tree-structured Filter Model (TFM). (© 2016, AIMS)

```

1: procedure DETECT_FACES(img)                                ▷ img is the face image query

     $VJ\_threshold \leftarrow 3$ ;                                ▷ minimum 3 merging bounding boxes
     $TFM\_threshold \leftarrow -1.1$ ;
     $VJ\_faces \leftarrow []$ 
     $final\_faces \leftarrow []$ 

2:    $VJ\_faces \leftarrow VJ(img, VJ\_threshold)$                 ▷ detect faces with VJ approach
3:   if  $VJ\_faces$  is not empty then
4:      $n \leftarrow length(VJ\_faces)$                             ▷  $n$  is the number of faces detected
5:     for  $i = 1$  to  $n$  do
6:        $face \leftarrow crop(VJ\_faces(i))$                       ▷ crop only the face region
7:        $face \leftarrow rescale(face, 40, 40)$                   ▷ resize the face into 40x40 pixels
8:        $score \leftarrow TFM(face)$                             ▷ verify the face with TFM
9:       if  $score \geq TFM\_threshold$  then
10:         $final\_faces \leftarrow final\_faces + face$           ▷ accumulate faces that pass the
        TFM threshold
11:      end if
12:    end for
13:  else
14:    print "no face detected in this image."
15:  end if
16:  return  $final\_faces$ 
17: end procedure

```

5.1.1 Model Training

The source code to train the TFM is publicly available (Zhu and Ramanan, 2012b). The TFM is almost similar to the MR-14 model with the same training dataset (Section 4.1.2) which are frontal faces of 112 subjects from AR database (Martínez and Benavente, 1998) scaled down to 10% resolution scale level (face size $\approx 30 \times 30$) along with the facial landmarks in ground truth (Ding and Martinez, 2010). However, there are three distinct changes compared to the MR-14 model. First, we only choose 12 landmarks to represent facial components as an indication of face presence. These landmarks consist of 2 eyebrow centres, 2 eye centres, 1 nose tip, 2 mouth corners, and 5 landmarks along the jawline. Second, we use less variation of facial expressions. Only neutral and scream expression are involved in the training (2 expressions \times 112 faces = 224 training face images). Lastly, in addition to 1218 images from INRIA dataset (Dalal and Triggs, 2005), random 1650 small-scale non-face images were added to negative training image set to further improve its performance to distinguish between faces and non-faces. The visualization of TFM and the corresponding tree structure can be seen in Figure 5.2.

5.1.2 Experiment Setup

We conducted the experiments on FDDB database (Jain and Learned-Miller, 2010). As mentioned in (Section 2.5.2.1), this database is suitable for evaluating face detection approaches with its well-made face ground truth and evaluation framework. Since our scope is on frontal faces only, 1535 images were manually chosen containing 2130 frontal/near-frontal faces (Figure 5.3).

The proposed combination of the Viola Jones & TFM is compared with the Viola Jones itself and Share-146 model. We did not integrate the MR models in this experiment since we want to measure the performance of proposed TFM. The Share-146 by Zhu and Ramanan (2012a) was once again chosen because of its capability to detect smaller faces (down to approximately 80×80). Since the Share-146 contains 13 models on various poses, we speculated that it might have an impact of detecting more faces, thus with possibility of increasing the chance of false detection on cluttered background. Therefore, we performed the evaluation of Share-146 on two scenarios: **1)** the Share-146 with all 13 models and **2)** the Share-146 with frontal face model only. We expect that the Share-146 with only a single frontal face model will perform with less false positives.

Evaluation on FDDB is based on the Receiver Operating Characteristic (ROC) curve to

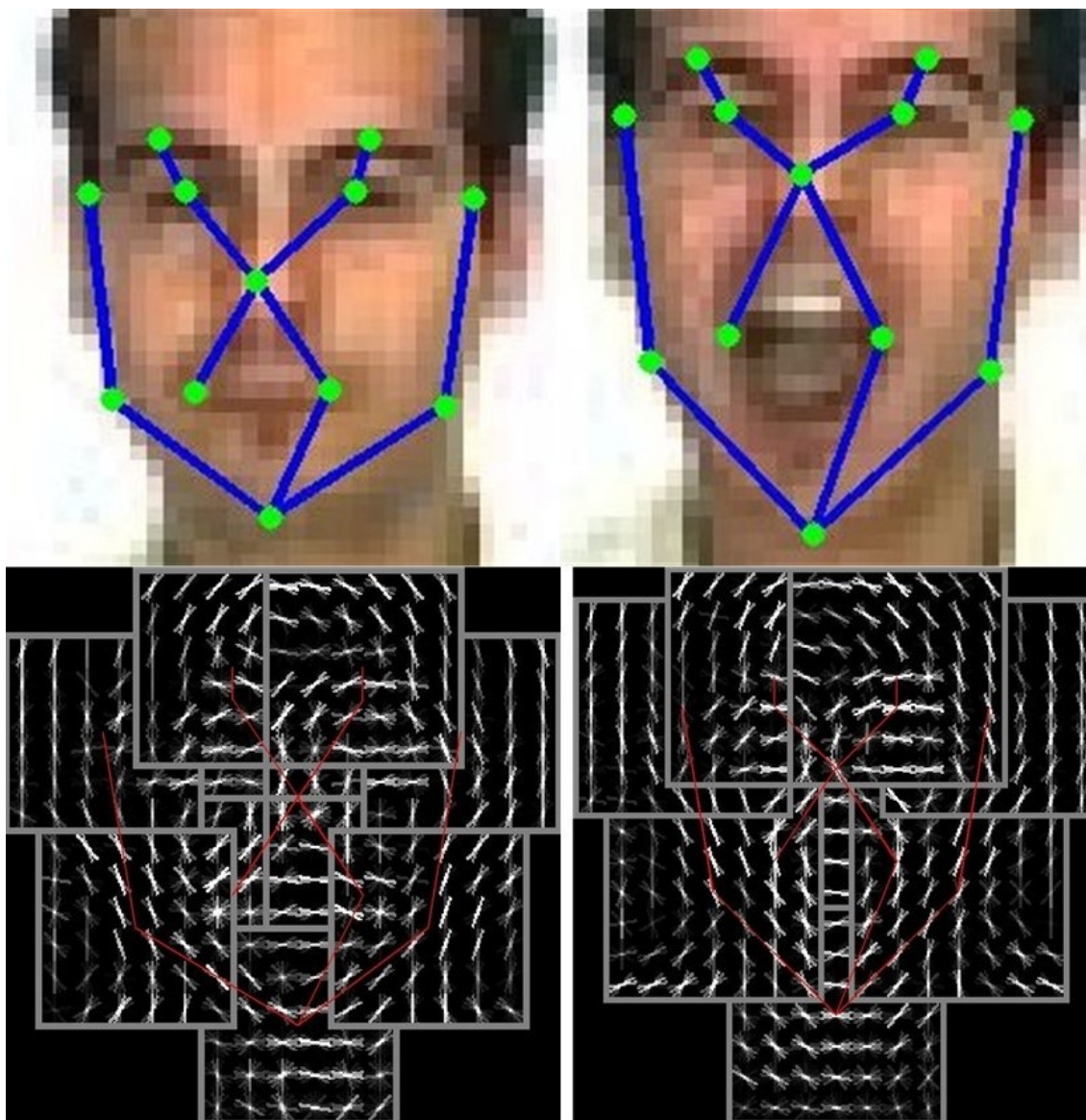


Figure 5.2: Visualization of TFM on neutral (**left**) and scream (**right**) facial expressions.
 (© 2014, IEEE. 2016, AIMS)



Figure 5.3: Some chosen frontal faces from FDDB database. (© 2016, AIMS)

plot the relation between true positive rate and false positive rate on various scores. We sort the true positive rate and false positive rate from the highest score to the lowest score for each technique. The score for the TFM and Share-146 is defined as the feature matching score described in the source code (Zhu and Ramanan, 2012b). We chose -1.1 and -0.75 as the minimum threshold for the TFM and Share-146 respectively. On the other hand, we define the score of the Viola Jones detector based on the amount of overlapping detection sub-windows merged together as a single face subwindow (Mat, 2012).

Ground truth comparison in FDDB is based on two types of metrics: discrete score and continuous score. The continuous score depends on the degree of match between the detected sub-window and ground truth which is defined as ratio between intersecting and joined region. For discrete score, a face is considered detected if the intersecting area is greater than 50% of the joined area. We only emphasize on discrete score since our main concern is on the presence of faces. We passed the detected rectangle sub-windows from each approach into the source code provided. The sub-windows from the Viola Jones detector were expanded approximately 30% to each side to ensure the faces were sufficiently covered for facial landmarking phase. Meanwhile, since the Share-146 can also detect the landmarks simultaneously, the detection sub-window is based on the border of the landmarks with the nose tip at the center to cover forehead region better. The examples can be seen in Figure 5.4.

5.1.3 Experiment Results

The ROC curves for all these face detectors are shown in Figure 5.5. One can see that the Viola Jones detector (VJ) achieves the highest detection rate, but also with the highest rate of false detection. This performance serves as the baseline performance. The Share-146 model is able to detect with significantly fewer false positives with a slight reduction



Figure 5.4: **(Left)** An example of query. **(Top right)** A face detected by Viola Jones and expanded prior to filtering by TFM. The subwindow is expanded to ensure sufficient coverage of the whole face for the facial landmarking phase. **(Bottom right)** A subwindow of a face detected by Share-146 model. It is cropped based on the edge of landmarks and nose tip as a central part to include forehead region.

on true detection rate. As we expected, involving all 13 pose face models in the Share-146 makes it more susceptible to false positives. By focusing only on a frontal face model, the Share-146 can reduce the false positives even more with a small drop in detection rate. However, it can be seen clearly that the combination of the Viola Jones and our proposed TFM outperforms all other approaches by detecting the least amount of false positives while maintaining high detection rate with only a slight reduction. Some examples on the TFM removing false positives can be seen in Figure 5.7.

In addition to the ROC curve, we also measured the average processing time for each approach. We only consider a single frontal face model for the Share-146. The result is plotted in Figure 5.6. The Share-146 considerably requires much more processing time since it has to scan the whole image to simultaneously fit the facial landmarks. Since the TFM only focuses on face candidate regions passed by the Viola Jones detector, it does not give too much burden in the processing time.

5.2 Facial Landmarking System

After successful experiments on the TFM, we integrated it with the MR models to form a complete facial landmarking system as shown in Figure 5.8. This integrated system is now able to perform facial landmarking on multiple frontal faces in various resolutions captured in uncontrolled environment.

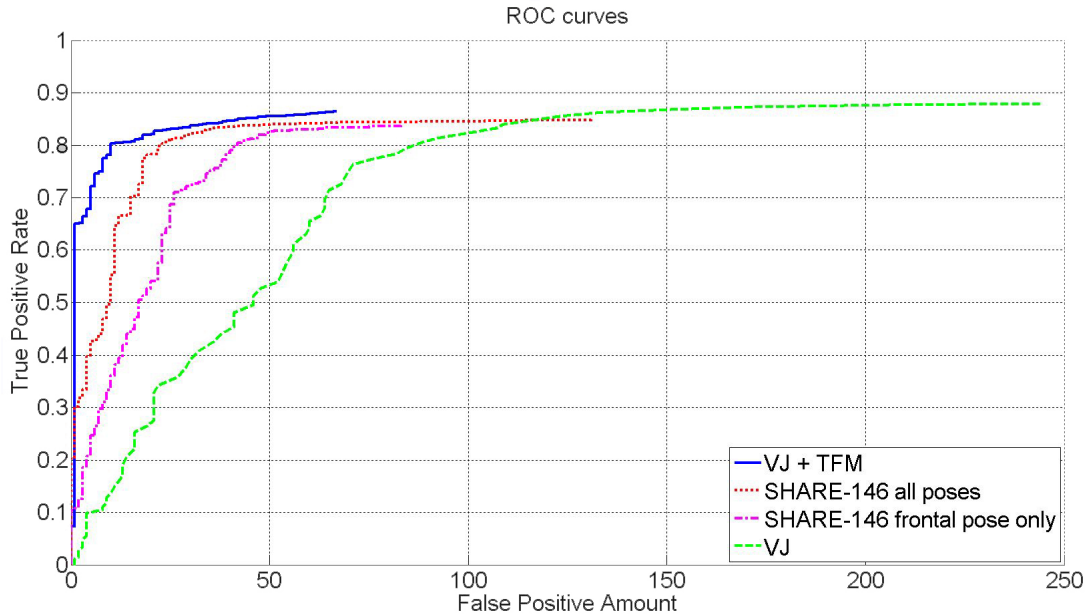


Figure 5.5: ROC (Receiver Operating Characteristic) Curve for Various Face Detectors.

5.2.1 Comparison with the Share-146

AFLW database (Koestinger *et al.*, 2011a) was chosen as testing dataset for our proposed facial landmarking system. As described in Section 2.5.2.2, this dataset contains a large amount of random face images suitable for real-life applications. In order to test on frontal/near-frontal faces only, we manually selected 200 images containing 687 faces. Most of these images contain multiple faces, some even in various resolutions (Figure 5.9).

Initially, we want to evaluate the performance based on landmark accuracy. Unfortunately, the ground truth provided in AFLW database lacks of accuracy and also has no sufficient quantity which makes it difficult to compare with (Figure 5.10). Alternatively, we manually count the number of faces which have been landmarked properly and false detection. The results are shown in Table 5.1. It shows that our proposed system outperforms the Share-146 model by a large margin on both true and false detection rate. Some visual results for comparisons can be seen in Figure 5.11.

5.2.2 Speed Comparison

Finally, we conducted an experiment to measure the growth of computational time as the image size gets larger. The purpose of this experiment is to show the advantage of having

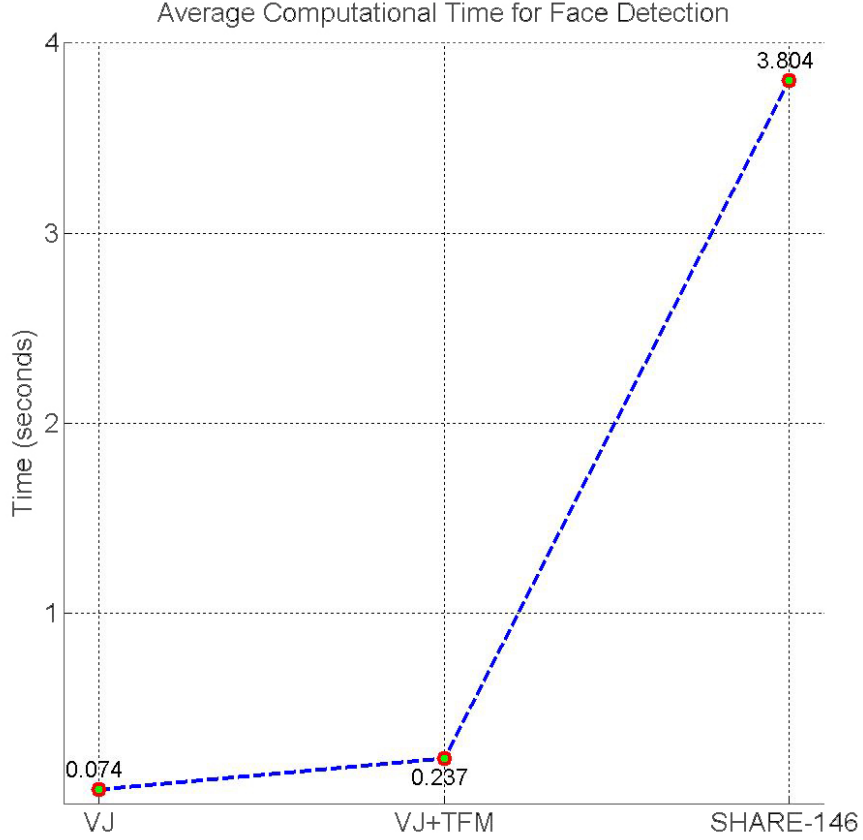


Figure 5.6: Time Comparison between VJ, VJ+TFM, and SHARE-146.

face detection (VJ + TFM) prior to facial landmarking approach. By focusing on the promising face candidates, processing time will be much less for the small size of the facial regions. On the other hand, even though the Share-146 was designed for better efficiency (Zhu and Ramanan, 2012a), it still needs to scan the whole image with extensive time cost.

In this experiment, we compare the Share-146 (frontal face model only) and our proposed system with the MR-36 (36 x 4 expressions = 144) since they both contain similar amount of facial landmarks. They were tested on two different scenarios (Figure 5.12). The first

Table 5.1: True Positive and False Positive on AFLW Database from the SHARE-146 Model (all 13 poses and single pose) and MR Models. (© 2016, AIMS)

	True Positive	False Positive
SHARE-146	473 (68.85%)	139
SHARE-146 (Frontal Only)	442 (64.34%)	7
VJ + TFM + MR	597 (86.90%)	2



Figure 5.7: **(Left)** The face candidates detected by Viola Jones detector. **(Right)** False positives are removed by our proposed TFM. (© 2016, AIMS)

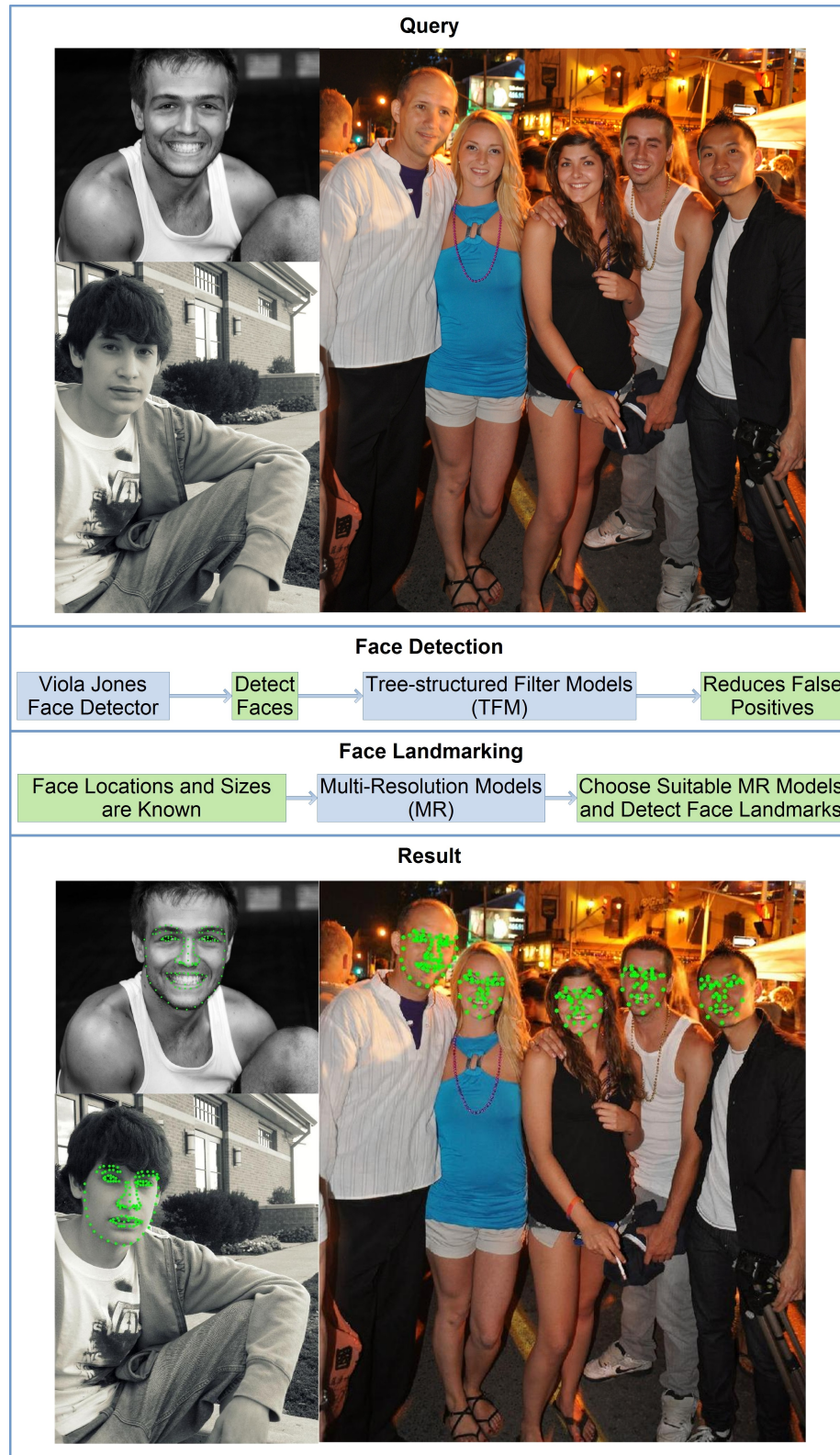


Figure 5.8: The integrated system combined from Viola Jones detector, TFM, and MR models. (© 2016, AIMS)



Figure 5.9: Chosen images from AFLW database. (© 2016, AIMS)



Figure 5.10: **(Left)** Some faces might not have landmarks ground truth. **(Right)** Some ground truth are not sufficiently accurate for comparison. (© 2016, AIMS)



Figure 5.11: Images on the left column are detected by Share-146 model (frontal model only), while the ones on the right column are detected by our proposed system. Share-146 miss the small faces and a false positive is detected on the background. ((© 2016, AIMS)

scenario is the case where the face occupies a small segment of the whole image (original size 579x389). On the other hand, the second scenario involves faces occupying large portion of the image (approximately 40% in this example) (original size 500x335). Both images were interpolated on 5 scale levels: 100% (original), 200%, 300%, 400%, and 500%.



Figure 5.12: **(Left)** Face region occupy a very small portion of the image. **(Right)** Faces occupy a large portion of the image (approximately 40%). (© 2016, AIMS)

We first observe the result on first scenario in Figure 5.13. As expected, the processing time of Share-146 escalates quickly from 7.7 to 202.5 seconds in 500% scale level. In comparison, our proposed system is not significantly affected by it.

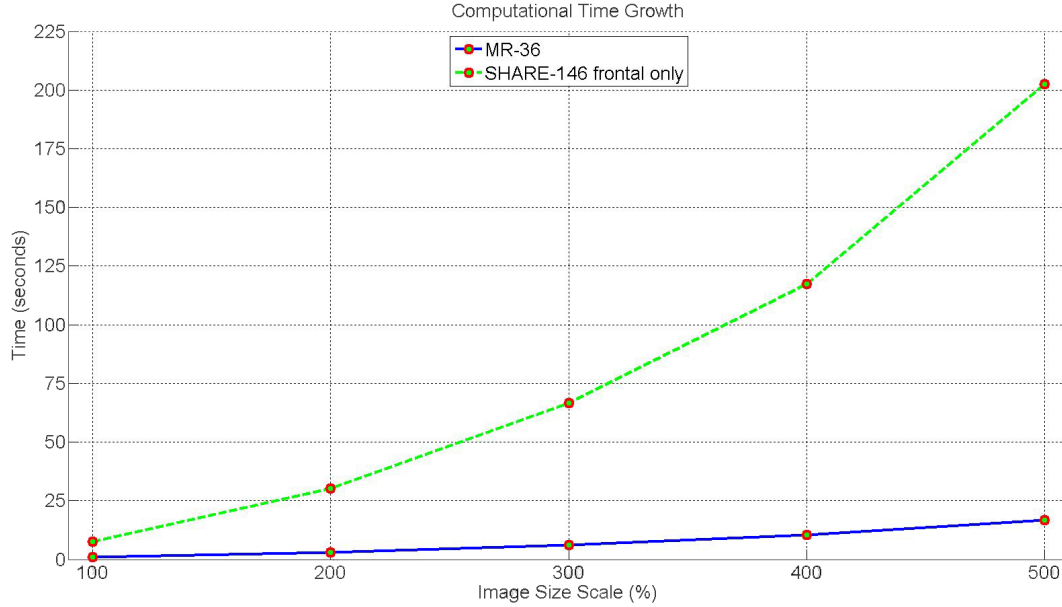


Figure 5.13: Speed comparison on a face on small segment of the image (**first scenario**). (© 2016, AIMS)

The second scenario is where the performance gap is much less. As shown in Figure 5.14,

our proposed system becomes significantly slow as the image gets larger even though it is still better than the Share-146. Since the face regions are large, scaling up the images still significantly increases the amount of data to be processed. Fortunately, this can be solved simply by fixing the size of the face regions in a compatible size with the corresponding MR models regardless the size of the image. For instance, the MR-36 can detect landmarks on faces with size at least 90x90. Despite its capability to handle much larger faces, it just leads to redundant processing. In this experiment, we fixed the size in a slightly larger resolution 150x150. As can be observed in Figure 5.15, this method significantly reduces the growth rate in any image resolution.

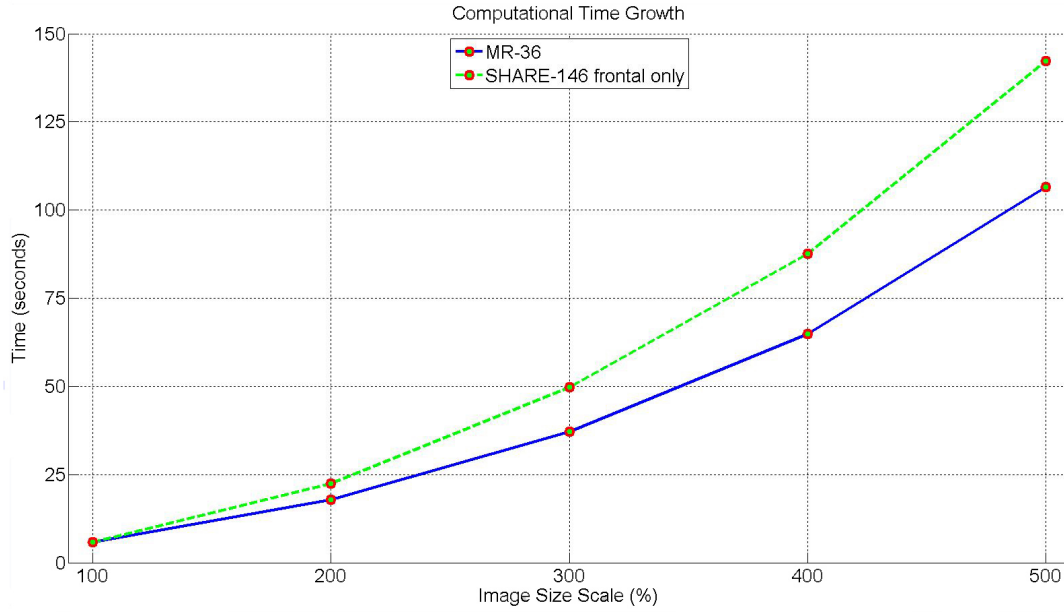


Figure 5.14: Speed comparison on faces on large segment of the image (**second scenario**).
(© 2016, AIMS)

5.3 Summary

In this chapter, we proposed a Tree-structured Filter Model (TFM) to act as a *filter* to discard as many false positives as possible from the Viola Jones face detector while preserving high rate of correct detections. The TFM was trained on low resolution faces (resolution $\approx 30 \times 30$) from AR database with restricted amount of landmarks and facial expressions. The chosen landmarks are 2 eyebrows, 2 eyes, 1 nose tip, 2 mouth corners, and 5 on jawline as to depict *intuitive description* of frontal human faces. It consists of only neutral and scream expressions. The reason for this restriction is to produce a lightweight model with low computation requirement. All face candidates detected by the

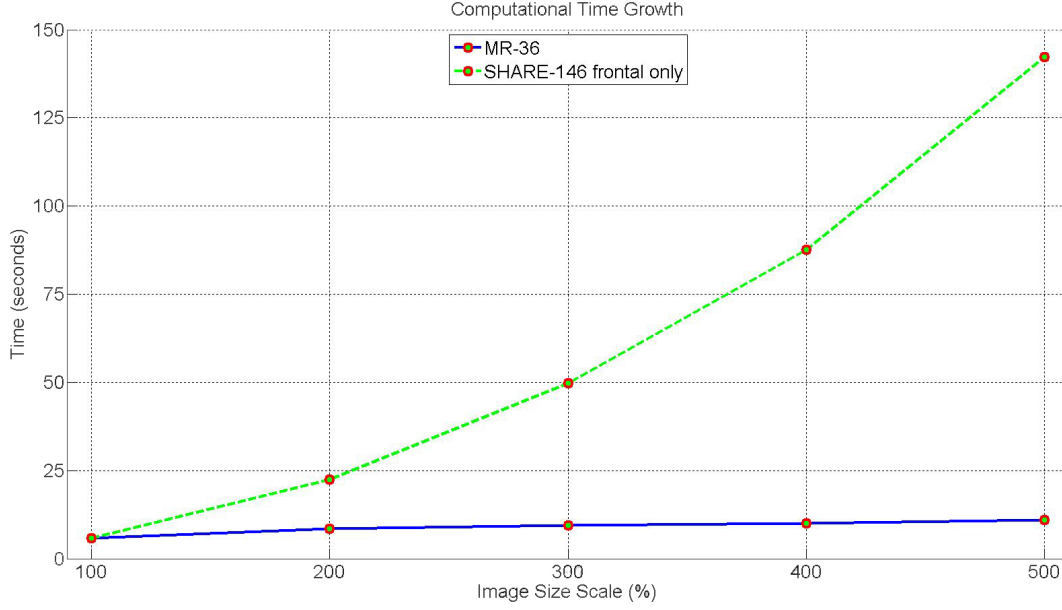


Figure 5.15: Speed comparison after scaling down faces to 150x150 on MR-36 models (**second scenario**). (© 2016, AIMS)

Viola Jones detector were scaled down to 40x40 to make it compatible with TFM since it was trained on small faces and simultaneously limit the amount of data to be processed for efficiency. Finally, the proposed TFM along with the Viola Jones detector complement MR models (Chapter 4) as an integrated facial landmarking system.

We conducted the experiments on two uncontrolled databases. We manually handpicked images containing only frontal/near-frontal face(s) due to the scope of our research. The first one is FDDB database which is suitable for evaluating face detection approaches. We compare the capability of TFM with the Viola Jones detector and Share-146 model. The result shows that combination of the Viola Jones detector and TFM outperforms other approaches with the lowest amount of false positives. For the next experiment, we evaluated the proposed integrated facial landmarking system in comparison with the Share-146 on AFLW database. The results clearly indicate a higher detection rate with lower false detections for the proposed system. Lastly, we conducted an experiment to show how the face pre-detection can reduce the processing time.

Chapter 6

Glasses Detection and Removal for Face Recognition and Verification

As the glasses/spectacles are widely worn for either fashion or visual problems (e.g short-sighted), the high usage of glasses is common in real world (Gao *et al.*, 2008). Therefore, the presence of glasses can be considered as a face semantic feature. For instance, Alattab and Kareem (2013) and Vaquero *et al.* (2009) have proposed semantic-based image retrieval involving glasses as one of the features. Since our previous experiments have shown that the concept of pictorial-tree-structured models can produce satisfactory performances in the context of *detection rate* and *landmarking* of the object of interest (frontal human faces for our cases), we would like to extend this usage into glasses as detection object. With these glasses models, we can detect the presence and location of a glasses on frontal faces and then manage to remove them in order to improve facial recognition performance. We believe this is necessary since the presence of glasses has a potential to negatively impact recognition proficiency (Righi *et al.*, 2012).

To the best of our knowledge, there are a few number of researches conducted on glasses-related applications in pattern recognition and computer vision community in the last two decades. Jiang *et al.* (2000) might be one of the earliest attempts to detect the presence of the glasses. Their approach is based on the level of intensity differences measurement surrounding the eyes. Their assumption is that it is high likely for glasses to have significantly different colour compared to facial skin, leading to a high level of intensity discontinuity around eyes, indicating its presence. Another approach proposed in the same period was proposed by Jing *et al.* (2000) via incorporating the Bayes rule on edge features extracted from Sobel filter. Furthermore, they attempted to remove the contour of the glasses by applying adaptive median filter. A few years later, Wu *et al.* (2004) adopted the idea of Markov-chain Monte Carlo technique for localizing the glasses segment and passing it through reconstruction process to remove the glasses for image synthesizing purpose. In spite of the significant performances on glasses detection and removal by these approaches, it is still limited to visual perception. There is no experiment conducted to measure the effect of their approaches on facial recognition.

However, there are other implementations with the aim of improving facial recognition rate. For instance, Wang *et al.* (2010) proposed an idea to localize glasses with Active Appearance Model (AAM) technique (Cootes *et al.*, 2001) and remove it via reconstruction process with PCA (Turk and Pentland, 1991). Despite a significant improvement made on the accuracy of facial recognition, there is no experiment to evaluate its glasses detection accuracy which determines whether a person is wearing a glasses or not. Another unique approach was proposed by Heo *et al.* (2004) where they combined the information from both visible features (pixel values) and thermal infrared (IR) images. This idea of utilizing thermal infrared was extended further by Wong and Zhao (2013) by attempting facial reconstruction on infrared space relying on the information around the eyes from the normal image. This novel way of including extra data from thermal infrared images shows a significant performance improvement, however it creates a restriction to have a specific device used for capturing thermal infrared images. We believe it is more preferable to focus only on colour/grey-scale images since they are more widely available.

Therefore, we intend to develop a complete autonomous glasses detection and landmarking which works on any frontal face images in this chapter. Furthermore, we remove the presence of the glasses based on the extracted landmarks with the aim to improve the facial classification performance. We achieved this by integrating our proposed glasses models with image reconstruction techniques: the NLCTV inpainting (Duan *et al.*, 2015) and SFDAE Deep Learning model (Pathirage *et al.*, 2015) (as mentioned in Section 2.3). We want to develop a system which can detect the presence and location of glasses automatically without assuming its existence (able to distinguish faces with and without glasses). Since it is difficult to find publicly available database specifically designed for glasses model training, we compiled various glasses and non-glasses images from CMU multiPIE (Gross *et al.*, 2010) as training dataset.

The robustness of our proposed glasses models is evaluated on various face databases. We apply our models on a large collection of face images with and without glasses. Afterwards, we evaluate the facial classification (recognition and verification) performance of the whole system based on three well-known classification techniques PCA (Turk and Pentland, 1991), LDA (Belhumeur *et al.*, 1997), and SRC (Wright *et al.*, 2009) as mentioned in Section 2.4.

The structure of this chapter is as follows. Section 6.1 describes the overview of our proposed glasses detection/landmarking and removal framework. It includes the training setup for the proposed glasses models along with the glasses images data we manually selected. Section 6.2 describes the setup of performance evaluations. We conduct experiments on *glasses detection rate* of our proposed glasses models and observe the impact of

glasses removal on facial classifications: *recognition* and *verification*. Lastly, the summary of this chapter is discussed in Section 6.3.

6.1 Framework

Our proposed system consists of two major parts: *glasses detection/landmarking* and *glasses removal/reconstruction*. In order to conduct the first part, we proposed a tree-structured glasses model as an alternative utility of face models by Zhu and Ramanan (2012a). We need to detect the presence of the glasses since attempting to reconstruct non-glasses faces is redundant or even negatively impacts facial recognition. Location information provided by the landmarks produced by these models are used to generate an image masking layer as a pre-processing stage prior to reconstructing glasses region. In the second part, we apply the image reconstruction approaches (the NLCTV inpainting (Duan *et al.*, 2015) and SFDAE Deep Learning model (Pathirage *et al.*, 2015)) described in Section 2.3 to remove glasses.

6.1.1 Face Alignment

Prior to glasses detection, we need to detect the presence and location of the face. It is essential to align the faces by ensuring equal face proportion and dimension for all query face images which is required to commence holistic face classification. In addition, this leads to an easier and more efficient glasses detection since we can restrict the search on the upper face region. The face alignment is done via a few basic transformations such as rotation, cropping, and resizing. All faces were scaled down to 360x320 with both eyes and mouth centres as parameters at specific locations. In this case, the chosen locations were inspired by face proportion in art done by MacTaggart (2000). In detail, we can assume that the origin coordinates are located in the top left corner of an image, we adjust the position of eyes at the proportion of $\frac{3}{10}$ and $\frac{7}{10}$ horizontally and $\frac{1}{3}$ vertically. The average of eye centres defines the horizontal position of mouth centre. Lastly, the vertical proportion for mouth centre is located at $\frac{10}{13}$. The visualization of our facial alignment proportion can be seen in Figure 6.1.

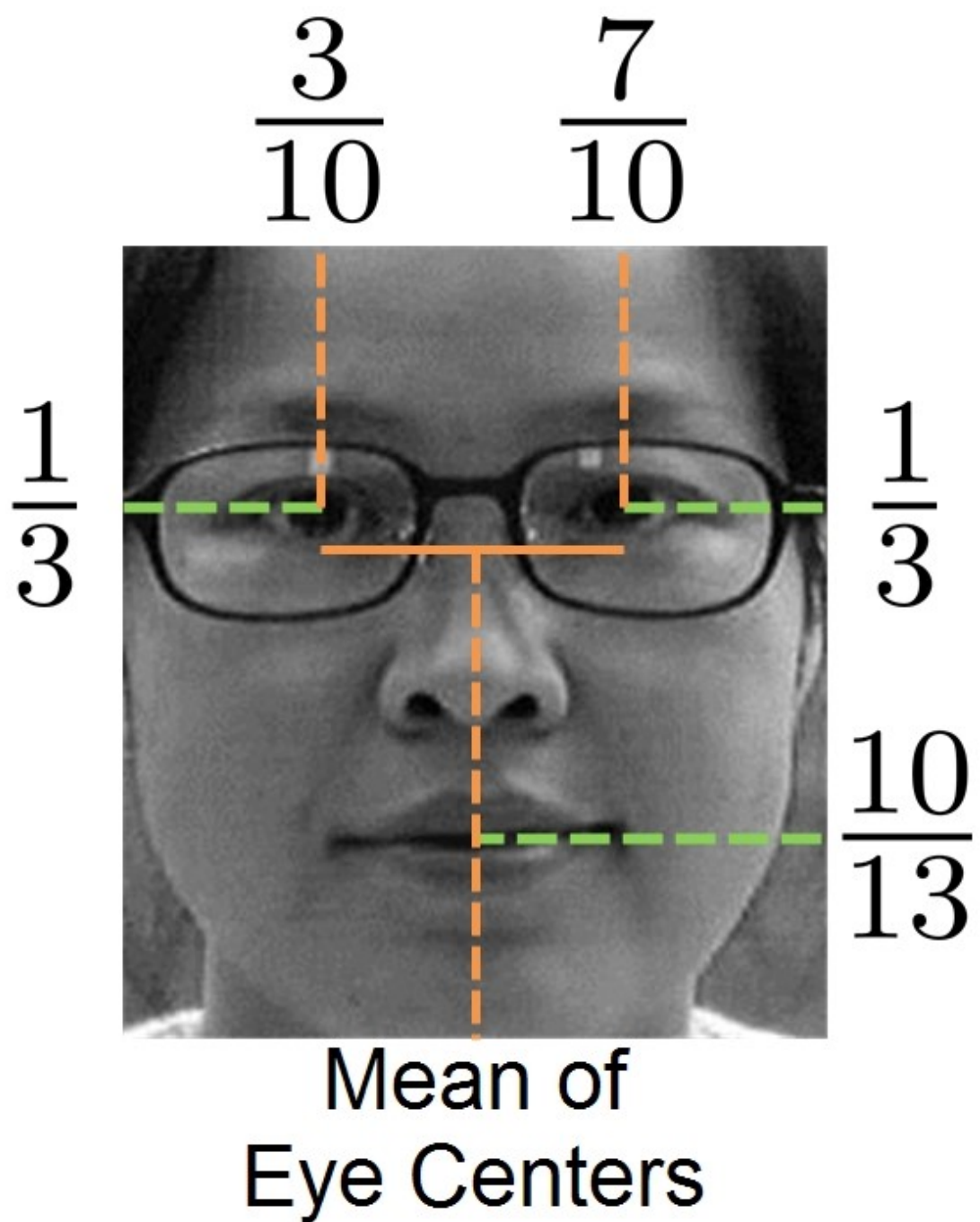


Figure 6.1: Aligned face based on the proportion of eye centres and mouth centre. The face is then scaled to 360x320. (© 2015, IEEE)

6.1.2 Glasses Model

As far as we know, there is no publicly available database specifically focused on glasses complemented with the corresponding landmarks in ground truth. Therefore, we have to manually choose some glasses-wearer face images ourselves and created the glasses landmarks. The images were extracted from CMU Multi-PIE database (Gross *et al.*, 2010). 100 neutral faces with glasses images were selected as the positive training set. This set contains 50 images with oval-shaped frame and other 50 with rectangle-shaped with round corner. We train two glasses models for both shapes. On the other hand, we did not use the same negative training set (the non-face images from INRIA database (Dalal and Triggs, 2005)) as in previous chapters' experiment. Since previous chapters emphasize on detecting/landmarking faces, it is reasonable to utilize the non-faces images to derive the false samples. However, the situation here is totally different for glasses. If we use the non-faces images as negative samples, the glasses models can not distinguish between glasses and non-glasses faces well which makes the whole framework not working properly. Instead, we selected 536 non-glasses neutral faces from the same database and cropped the region around the eyes as the negative training set. Due to the restriction of the publication of the faces, we can only show the examples of glasses we chose in Figure 6.2.



Figure 6.2: Two chosen glasses shapes: (**Top**) Oval (**Bottom**) Rectangle. (© 2015, IEEE)

Unlike facial landmarks, we could not find glasses landmarks in ground truth available for our model training. So, we had to manually create our own ground truth for those 100 positive training images. Each pair of glasses contains 39 landmarks with the distribution: 32 landmarks along both rims, 3 landmarks on the bridge between rims, and 4 landmarks on both hinges as shown in Figure 6.3. Furthermore, in order to have a more balanced

and consistent landmarking on the rims, we avoid pinpointing the landmarks in circular order directly. Instead, we did it in a *hierarchy* manner. We first appointed 4 landmarks on 4 orientations: leftmost, rightmost, top, and bottom corresponding to the center of the rim. This is to ensure a proper calibration of initial landmarks to adjust the balance of landmarks distribution. The next step is to insert a new landmark between two appointed landmarks to add 4 more landmarks in diagonal directions. Repeating this step one last time will eventually result in approximately uniform-distributed 16 landmarks. The process can be seen in Figure 6.4. With all these landmarks ground truth and the training dataset, we can train our tree-structured glasses models as shown in Figure 6.5.

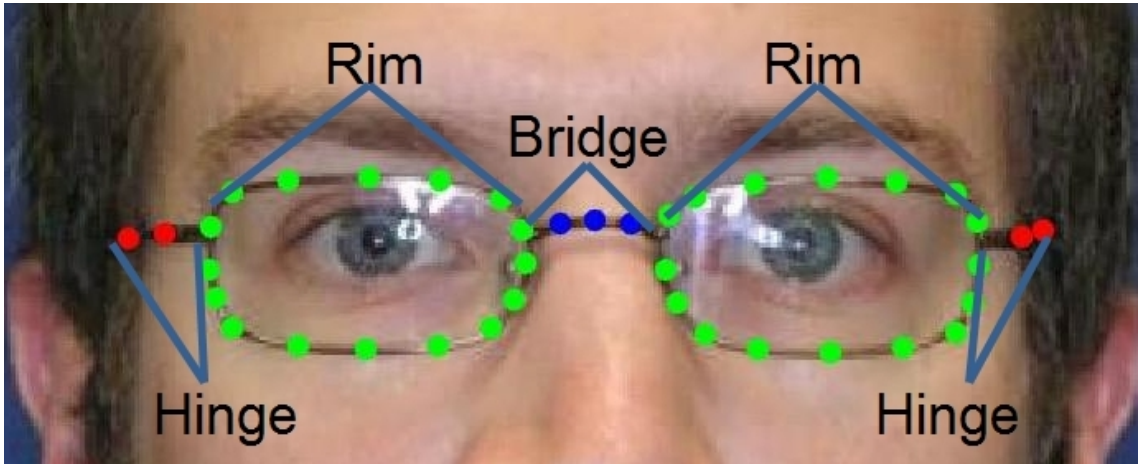


Figure 6.3: Our own created 39 glasses landmarks ground truth. (© 2015, IEEE)

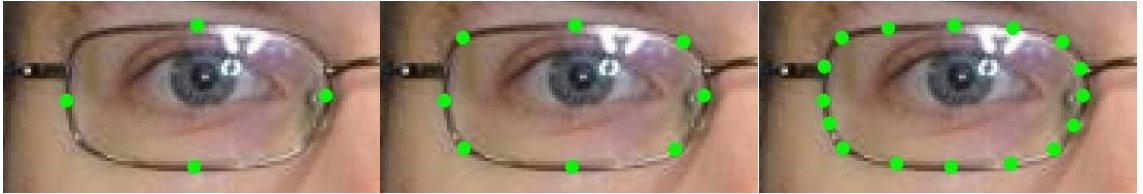


Figure 6.4: The process of appointing 16 landmarks on a rim. **(1)** 4 landmarks on the left, right, top, and bottom position. **(2)** A landmark is added in the middle of each pair of previous landmarks. **(3)** Repeat the last procedure once more to pinpoint the final landmarks. (© 2015, IEEE)

6.1.3 Masking

Based on the extracted landmarks, we create an additional layer of mask to indicate the location of glasses regions. This information is required in employing the NLCTV inpainting approach to reconstruct the glasses segments. The mask was derived by linking all the adjacent landmarks with linear interpolation (straight line) between each pair.

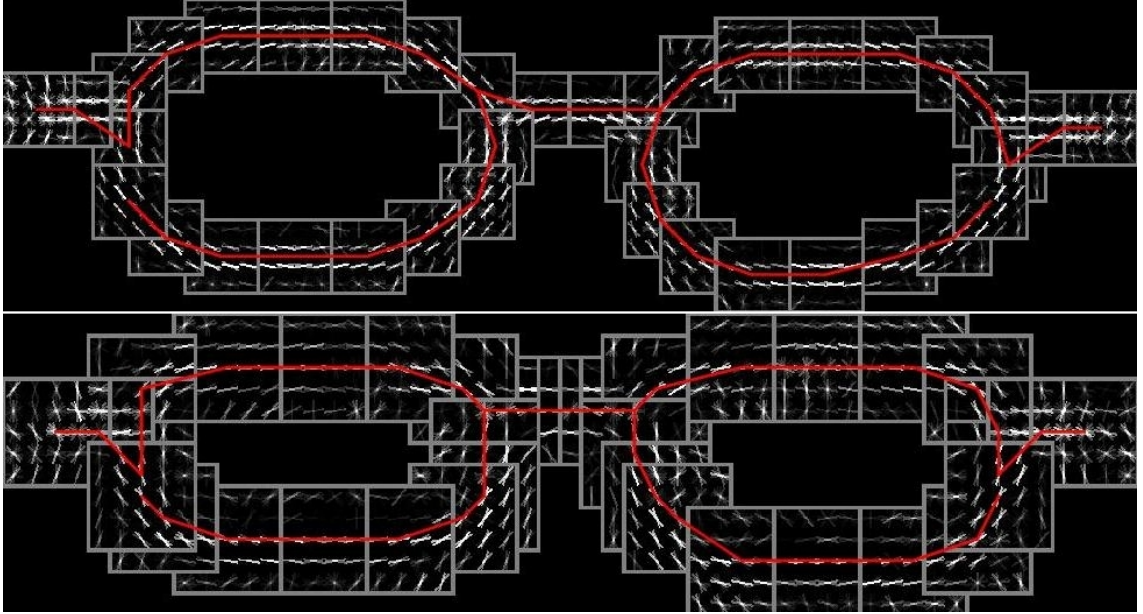


Figure 6.5: Our proposed glasses models. The first model is an oval-shaped glasses while the other one is a rectangle-shaped glasses. (© 2015, IEEE)

Since this will create a jagged surface, we smoothed the mask by adopting the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation (Kahaner *et al.*, 1989; Fritsch and Carlson, 1980; Moler, 2008). The visualization between the original mask and smoothed mask can be seen in Figure 6.6. Furthermore, in order to enhance the coverage around nose pad and bridge, we added another layer of a slightly wider mask. This additional layer also covers shadow on the lower rim for some cases. An example of a combined set of mask layers can be seen in Figure 6.7.

6.1.4 The Complete Framework

The glasses removal was conducted by recovering the region of interest via image reconstruction techniques as described in Section 2.3. We utilized two state-of-the-art image reconstruction techniques: the Non-Local Colour Total Variation (NLCTV) inpainting (Duan *et al.*, 2015) and Stacked Face De-noising Auto Encoders (SFDAE) Deep Learning model (Pathirage *et al.*, 2015). We arrange these approaches in a "cascade" structure starting with the NLCTV followed by SFDAE to act as a double-layered filters to remove the "noise" on the face images (Figure 6.8). In our case, the presence of glasses is considered as noise and thus should be removed. Since the NLCTV inpainting has removed most traces of the glasses, the de-noising phase will make it more accurate, for example, slight light reflection on the lenses can be removed.

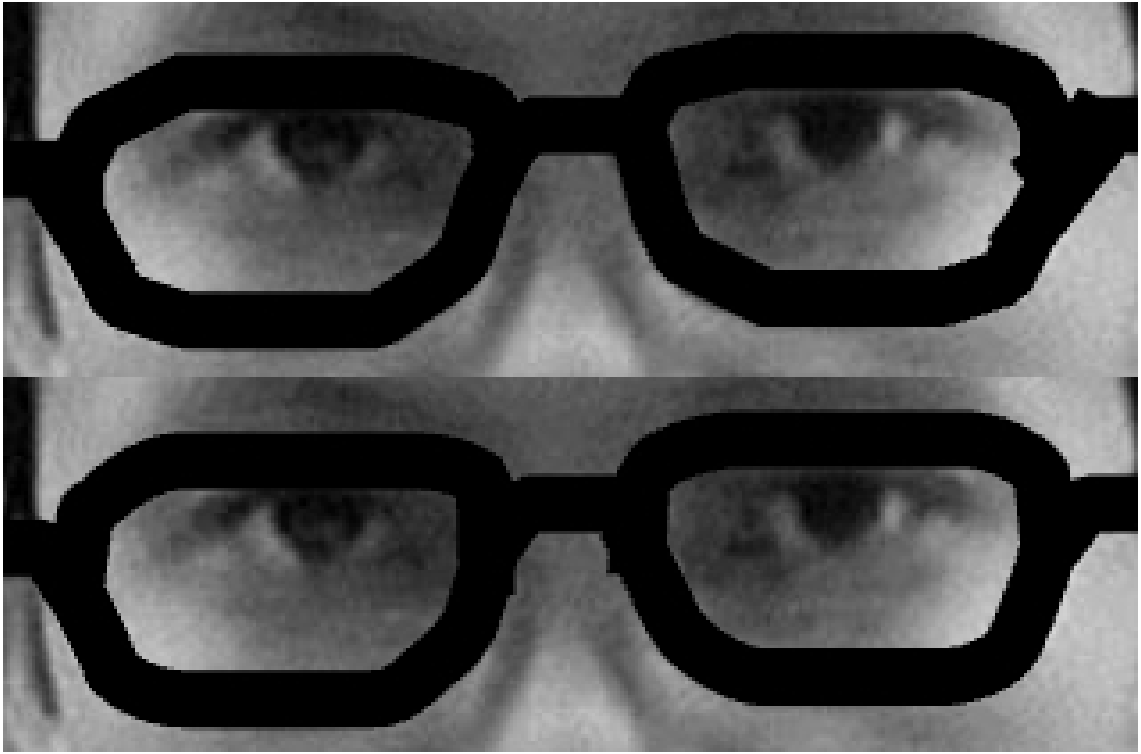


Figure 6.6: Masks derived by (**Top**) linear interpolation and (**Bottom**) PieceWise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation. (© 2015, IEEE)



Figure 6.7: (**Left**) First layer of mask covering all base parts of glasses. (**Middle**) Additional layer of mask to cover nose pad, bridge, and lower rim. (**Right**) Combination of both layers of mask. (© 2015, IEEE)

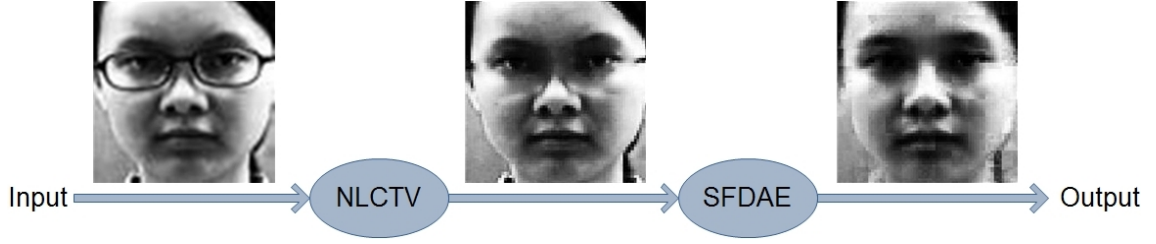


Figure 6.8: 2 state-of-the-art image reconstruction methods (NLCTV inpainting & SFDAE Deep Learning model) structured in a "cascade" manner to filter the presence of glasses consecutively.

The whole framework is summarized in Figure 6.9. It consists of four stages **(1)** Face alignment via face landmarking with the AR/MR-130 models as a pre-processing phase **(2)** Glasses presence detection and landmarking with our proposed glasses models **(3)** First phase of glasses removal process via reconstruction by NLCTV inpainting with the help from the mask as the NLCTV inpainting approach requires the boundary information. As observed in Figure 6.10, it demonstrates how it considers the glasses segment as noise and reconstructs it based on the surrounding skin texture. **(4)** The Second phase of glasses removal process via reconstruction by the SFDAE Deep Learning to remove last traces of glasses and slight light reflection as shown in Figure 6.11. Since the SFDAE is a patch-based approach, the face images from NLCTV inpainting have to be pre-processed. We first apply Histogram Equalization to normalize the illumination in the image. We then resize the image into 66x66 and divide the image with patch of size 6x6 resulting in 11x11 patches. Please be advised that the reconstructed image is only to visualize the result of glasses removal. We will use the low-dimensional features extracted from the de-noising layer f_2 (second hidden layer) for face classification as mentioned in Section 2.3.2.

6.2 Experiments

We conducted a few experiments to evaluate the performance on various stages of our proposed system. Basically, it evaluates on two major parts: *glasses detection/landmarking* and *impact of glasses removal/reconstruction* for face recognition. For the first part, we assessed the capability of our proposed glasses models to detect the presence of glasses on a face on various databases. The next stage was tested based on the improvement of facial *recognition* and *verification*.

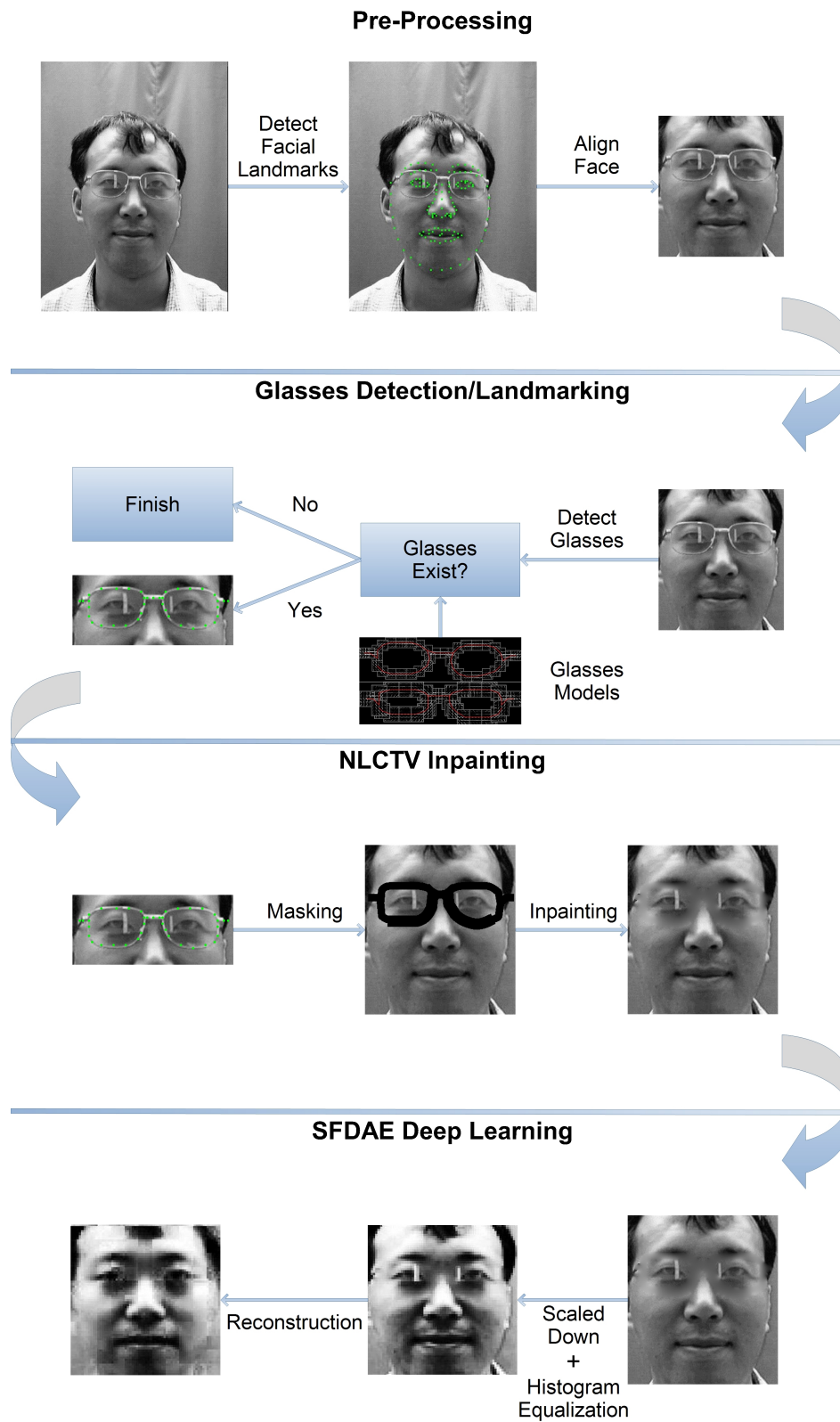


Figure 6.9: Our proposed glasses detection + removal system. (© 2015, IEEE)



Figure 6.10: This is the iterative process of NLCTV inpainting on glasses. The image on **top left** is the original face image with glasses. The next image shows the mask generated from the glasses landmarks which is then gradually reconstructed along with the glasses.



Figure 6.11: **(Left)** Aligned face images wearing glasses **(Middle)** Face images after NLCTV inpainting **(Right)** Face images after reconstruction via NLCTV inpainting + SFDAE Deep Learning model. All images in this example are contrast normalized through histogram equalization method and scaled down to 66x66. (© 2015, IEEE)

6.2.1 Glasses Detection/Landmarking

For this experiment, we selected frontal neutral faces from 6 face databases: CAS-PEAL-R1 (Gao *et al.*, 2008), CurtinFaces (Li *et al.*, 2013), AR (Martínez and Benavente, 1998), FEI (OLIVEIRA JR and Thomaz, 2006), PUT (Kasinski *et al.*, 2008), and BU-4D (Yin *et al.*, 2008). As summarized in Table 6.1, each database has different amount of face images with glasses. The selected dataset is mainly dominated by CAS-PEAL-R1 due to its large number of participants and particular session on various glasses. Initially, there are 438 subjects participating on wearing glasses. However, some of the images are affected by strong illumination on the lenses occluding significant part of the eyes. In order to avoid unfairness, we did a thorough selection resulting in 340 chosen subjects. As mentioned in Section 2.5.1.7, the accessories section consists of 3 different glasses. We only choose 2 images for each participant because the last one is sunglasses in some occasions which is not in the scope of our research.

Table 6.1: Information on chosen face images on various databases. (© 2015, IEEE)

	CAS-PEAL-R1	CurtinFaces	AR	FEI	PUT	BU-4D
People	340	52	136	200	100	101
No. Images	1020	104	136	200	100	101
Glasses	680	19	40	8	0	0
Non-Glasses	340	85	96	192	100	101

Since this is an experiment about glasses detection/landmarking, we cropped all the aligned query faces further to focus on the eye region. This is useful since we can avoid redundant computation and reduce the chance of false detection on the non-eye region. The range we chose is from row 61 to 200 and the whole 320 columns. All the cropped images were tested with matching score threshold -0.54 . The result of our proposed glasses model is summarized in Table 6.2. Our proposed glasses models achieved close to perfection in distinguishing face images with glasses or non-glasses. There is only one missed detection from FEI database as shown in Figure 6.12. Our investigation revealed that particular glasses is actually rimless, thus causing the level of intensity differences of the glasses edges too *faint*. This result is justifiable since our proposed glasses models rely on edge information (HOG features (Dalal and Triggs, 2005)) which makes it difficult to fit these glasses landmarks, hence producing a low matching score.

Table 6.2: Glasses detection rate on various databases. (© 2015, IEEE)

	True Positive	True Negative
CAS-PEAL-R1	680/680 (100%)	340/340 (100%)
CurtinFaces	19/19 (100%)	85/85 (100%)
AR	40/40 (100%)	96/96 (100%)
FEI	7/8 (87.5%)	192/192 (100%)
PUT	0/0 (N/A)	100/100 (100%)
BU-4D	0/0 (N/A)	101/101 (100%)



Figure 6.12: Since this is a rimless glasses, the edge features are too *faint* to consider it as a glasses. (© 2015, IEEE)

6.2.2 Glasses Removal/Reconstruction

The second part of our experiment is to evaluate the impact of glasses removal on face classification. More specifically, we performed facial recognition and verification. We only use the images from CAS-PEAL-R1 database for training, gallery, and testing set since it contains the largest amount of faces wearing glasses among other databases.

6.2.2.1 Inpainting

Prior to conducting facial classification evaluations, we did an evaluation on the NLCTV inpainting. Even though Figure 6.10 has shown how NLCTV inpainting can remove the glasses *visually* (image synthesis), we did not measure the result *numerically* for analysis. Therefore, we conducted an experiment to measure the impact of NLCTV inpainting in reducing the gap between the reconstructed faces and original non-glasses faces.

In this experiment, we choose 340 frontal non-glasses faces from CAS-PEAL-R1 explained in Section 6.2.1 as the gallery set. However, we did not use the set of faces with glasses since there ought to be a slight difference on various face parts despite being compared with the same subject. Since we want to measure the difference only from the glasses, these images can not be used. Instead, we created synthetic face images via incorporating additional layer of glasses on top of the face. We extracted two types of glasses: *thin silver* and *thick dark* through image editing software Photoshop and placed them on the eye region for each subject. The process of producing these synthetic images can be done automatically in face alignment (Section 6.1.1). The examples can be seen in Figure 6.13. In such a way, we can ensure the difference only comes from the glasses for accurate measurement.



Figure 6.13: **(Left)** Original image without glasses **(Middle)** First synthetic data with thin silver glasses **(Right)** Second synthetic data with thick dark glasses. (© 2015, IEEE)

We applied our proposed glasses models to both types of glasses and removed them with the NLCTV inpainting. We then measured the mean of l_2 -norm distance (Euclidean) between the synthetic data (both glasses and inpainted) and the original faces. The result is summarized in Table 6.3. It can be observed that inpainted glasses reduced the distance by approximately **42.28%** and **61.94%** for thin and thick glasses respectively. With this performance, we believe this will bring a positive impact towards face classification and makes it easier for reconstruction process with the SFDAE model.

Table 6.3: Average Euclidean distance between the synthetic data and neutral frontal face images. (© 2015, IEEE)

	Glasses	Inpainted
Thin	7409.41	4276.38
Thick	14764.35	5619.91

6.2.2.2 Face Recognition

We use the same collection of face images of 340 subjects from CAS-PEAL-R1 as experiments in Section 6.2.1. We conducted this experiment with a cross-identity setup. It means that training image set and gallery/testing image set will not share the same subject. Training set consists of 4 non-glasses images including neutral face per subject to train the transformation function of SFDAE. The non-neutral facial expressions are considered as 'noisy' faces, and we want to train the SFDAE model to reconstruct them into neutral faces via supervised learning. The trained model is used to attempt further reconstruction to remove the remaining traces of glasses after inpainting. In total, we choose 98 subjects in this set. On the other hand, the testing involves one neutral face image as the gallery and two glasses images as the query from each identity for the remaining 242 subjects. The illustration of the experiment setup can be seen in Figure 6.14. We investigated the results on three scenarios: face with glasses, inpainted glasses (NLCTV), and reconstructed glasses (NLCTV + SFDAE).

As mentioned in Section 2.4, we utilized 3 well-known linear classifier approaches to measure facial recognition rate: PCA (Turk and Pentland, 1991), LDA (Belhumeur *et al.*, 1997), and SRC (Wright *et al.*, 2009). The result is summarized in Figure 6.15. As can be expected, faces with the presence of glasses achieve the lowest performance. Since glasses add unnecessary noises, it disrupts the classification process. Inpainted glasses appear to provide slight improvement towards recognition rate. Our observation suggests two possibilities for this result. First, due to the restricted availability of data, we can only use CAS-PEAL-R1 which contains only **grey-scale** images. However, the NLCTV inpainting is able to reconstruct the image texture based on the color information, its full potential could not be utilized in this case. Second, the proportion of the inpainted area compared to the size of the whole face is relatively small. Even though the result is better, the changes only affect local parts of the face (around eyes). This why we added another layer of glasses filter via the SFDAE model. Its de-noising process covers the whole face including glasses regions. In addition, since the NLCTV has removed most of the glasses segments, it makes it easier for SFDAE to de-noise the remaining traces of the glasses and slight lens reflections. The significant improvement is achieved with this proposed scheme. The combination of NLCTV and SFDAE **reduces the error rate** by approximately **50%**, **52.25%**, and **57.09%** for PCA, LDA, and SRC respectively.






Training		
Gallery		
Testing	Glasses	
	NLCTV	
	NLCTV SFDAE	

Figure 6.14: Illustration of experiment setup for our cross-identity testing.

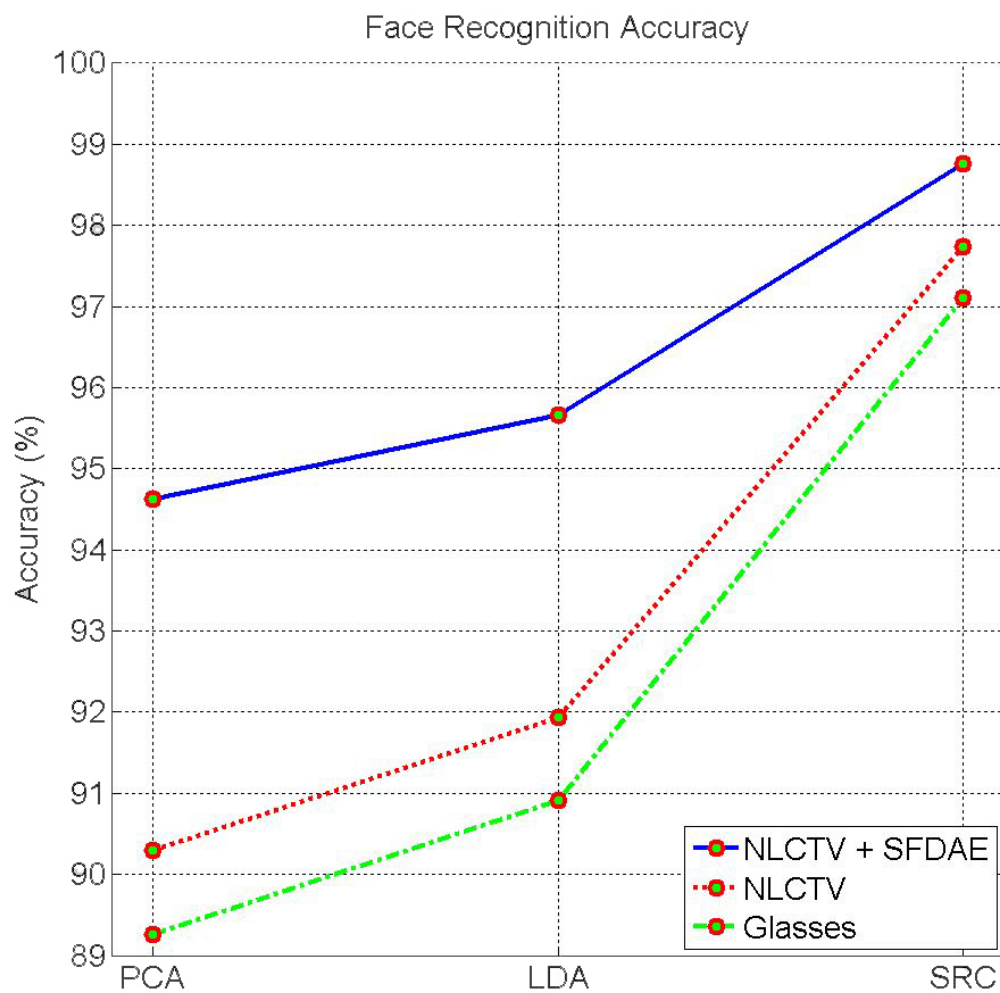


Figure 6.15: Facial recognition with classification approaches PCA, LDA, and SRC. This result proves that removing presence of glasses improves the facial recognition rate.

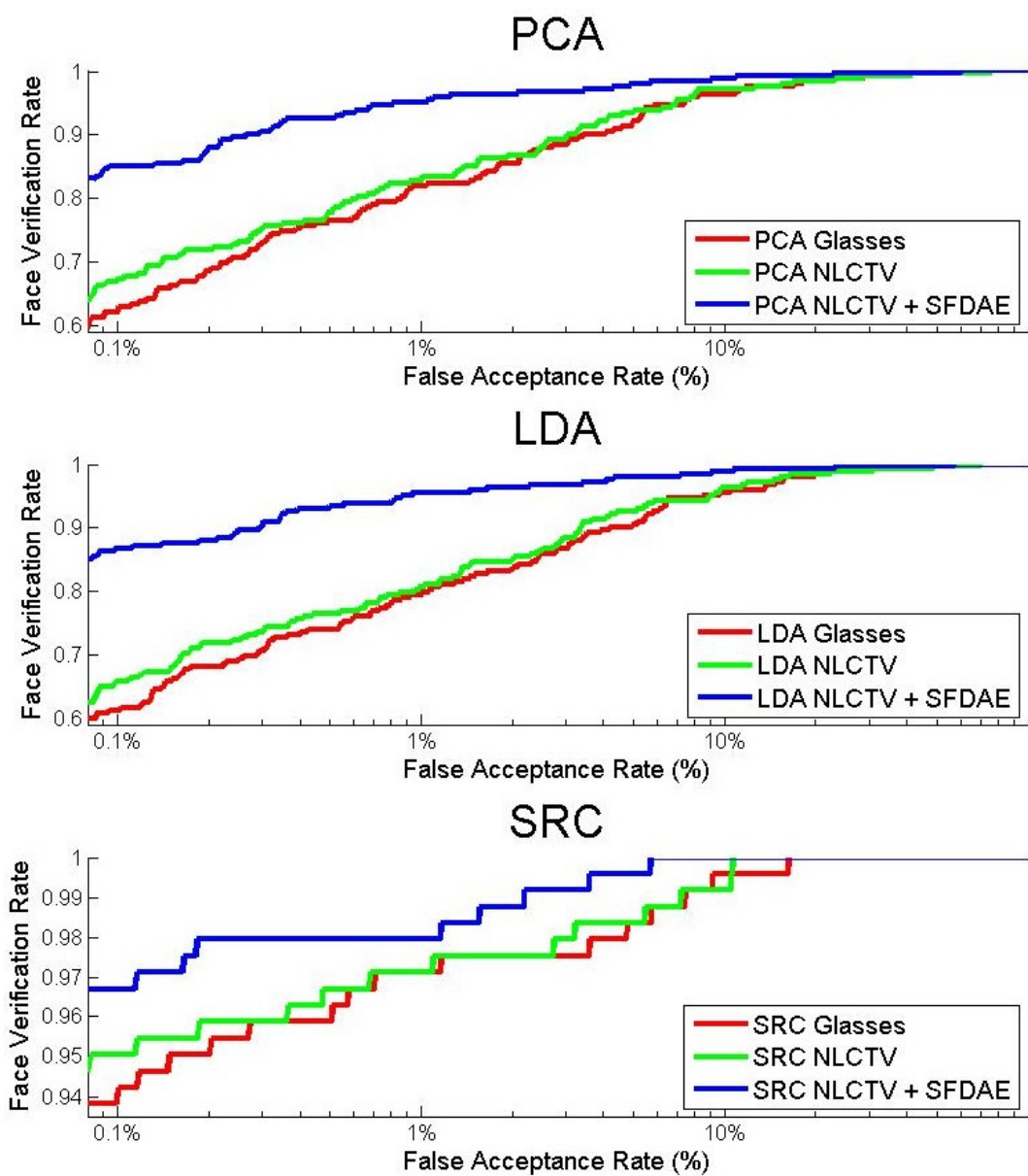


Figure 6.16: ROC curves on thin glasses.

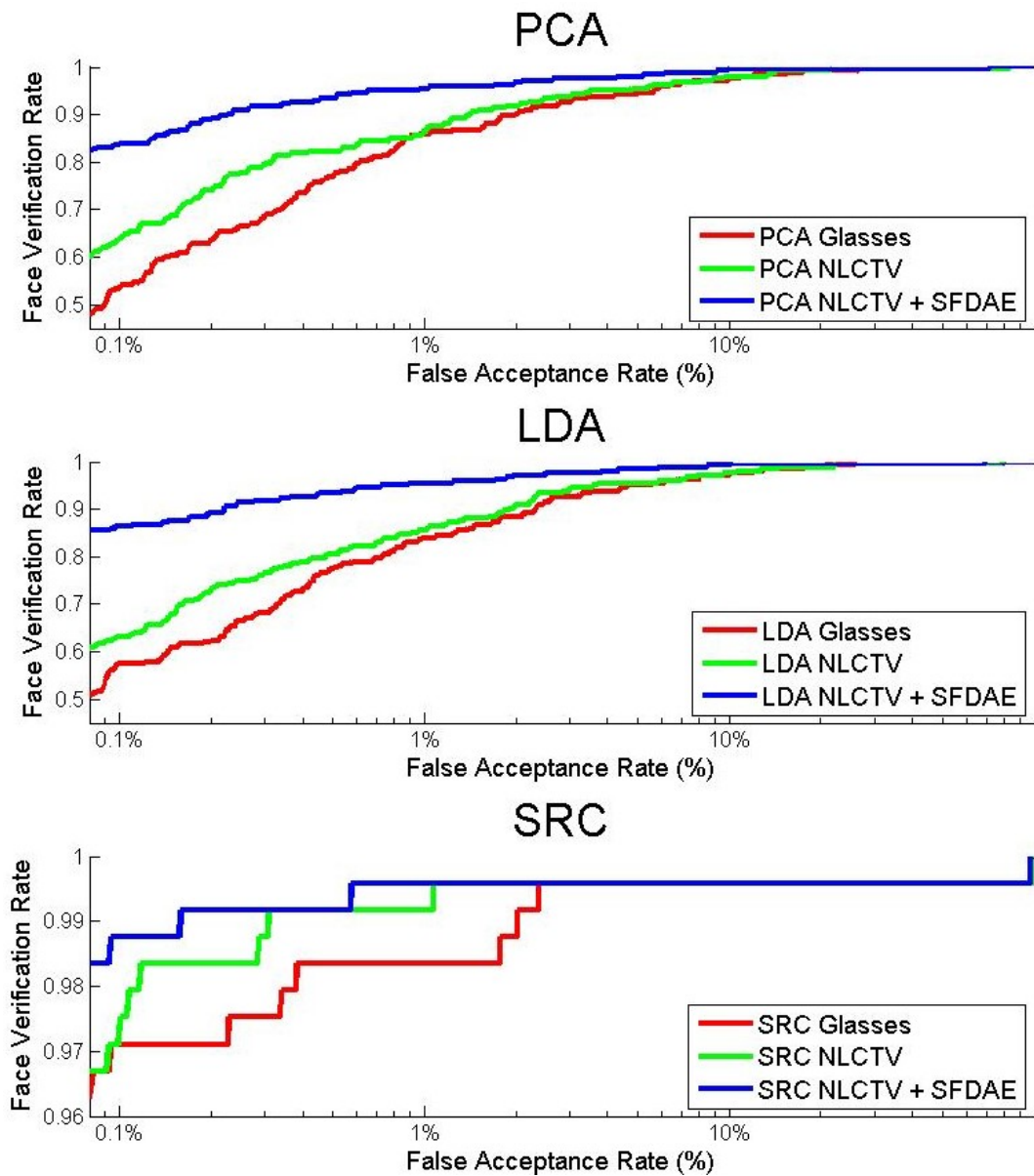


Figure 6.17: ROC curves on thick glasses.

6.2.2.3 Face Verification

The next experiment is based on the accuracy improvement on face verification. This is different from facial recognition where a query face is compared to a set of gallery images and choose the one with the highest matching score. Instead, face verification is a one-to-one face matching on which the decision is made based on a threshold of the score. The ROC curve is then generated on various thresholds. Choosing the best threshold is not a simple task due to the trade off between the true and false acceptance rate. It is widely accepted to only consider threshold with 0.1% False Acceptance Rate (FAR).

We first learned a transformation function with SFDAE model by using the same 98 training subjects images and evaluated it on 242 testing subjects (*thin* and *thick* glasses) same as the previous experiment. However, the testing setup is now different due to face verification's one-to-one matching nature. Neutral face from each subject can be paired with other 242 faces wearing glasses. This create a single correct pair and 241 false pairs for each participant. In total, we have 242 true matches and $242 * 241 = 58,322$ false matches from each scenario.

The test was conducted on two scenarios: thin and thick glasses. For each scenario, we classified the faces with PCA, LDA, and SRC. We compared the verification performance between the original glasses-wearer faces, inpainted faces (NLCTV) and reconstructed faces (NLCTV + SFDAE) images. The ROC curves are available in Figure 6.16 and 6.17. The verification rate at 0.1% False Acceptance Rate (FAR) is summarized in Table 6.4. As expected, the verification performance is significantly improved following the glasses removal process. The improvements are especially distinct with PCA and LDA.

Table 6.4: Face verification rate at 0.1% False Acceptance Rate (FAR) before and after glasses removal.

Classification	Thin Glasses			Thick Glasses		
	Glasses	NLCTV	NLCTV + SFDAE	Glasses	NLCTV	NLCTV + SFDAE
PCA	62.81	67.36	85.12	53.72	64.05	83.88
LDA	61.16	65.70	86.78	57.44	63.22	86.36
SRC	94.21	95.04	96.69	97.11	97.52	98.76

6.3 Summary

In this chapter, we proposed an automatic integrated glasses detection/landmarking and removal system for improving facial classification performance. This framework consists

of two major parts: *glasses detection/landmarking* and *glasses removal*. We proposed glasses models as an alternative concept of pictorial-tree-structured face models by (Zhu and Ramanan, 2012a). We proposed oval-shaped and rectangle-shaped glasses models trained from 100 face with glasses images manually chosen from CMU multiPIE database along with our own 39 landmarks ground truth. Furthermore, in order to improve its robustness to distinguish between faces with glasses and non-glasses, we manually selected 536 cropped eye regions from non-glasses face images as negative samples. The landmarks extracted via these models are used to localize glasses segments through masking process for glasses removal phase. We integrated our proposed glasses models with two image reconstruction techniques: *NLCTV inpainting* and *SFDAE Deep Learning model* as a double-layered filter to remove the presence of glasses. The experiment results reveal the high performance of our proposed glasses models on detecting the presence of glasses on various face databases. Further experiments demonstrate positive impacts of removing glasses towards both facial recognition and verification.

Chapter 7

Face Retrieval based on Semantic Features via Face Landmarking

As it is known that the works based on semantic representations are common on retrieving documents (e.g text documents) (Mangold, 2007) or contents inside the images (Liu *et al.*, 2007). Face-related applications are not exceptional on this field (Karczmarek *et al.*, 2015). For instance, Wang *et al.* (2016) proposed an approach to classify Chinese ethnic groups from facial semantic features. Another example is the software used by law enforcement EvoFIT (Frowd *et al.*, 2004) to identify criminal suspects by creating composite sketch based on the descriptions by the crime witness. For the scenario of face images retrieval, one of the classic approaches is conducted by Gudivada *et al.* (1993) by deriving facial semantic attributes via Personal Construct Theory (PCT) (Kelly, 1955, 1969). However, it needs a domain expert to do manual iterative selection of the semantic attributes. Another approach is accomplished by Sridharan (2006). The author proposed a framework to extract the features of facial components automatically (e.g probabilistic approach and polygon fitting). However, the semantic information are quite simple such as the height and width of the facial components. We want to include more sophisticated features such as the geometrical shape of facial components as done by Conilione and Wang (2012). Unfortunately, despite the complex semantic features, Conilione and Wang (2012) had to manually created all the facial landmarks for all the face images which is time-consuming for registering semantic membership degree and alignment.

These limitations motivate us to design a face images retrieval system which can perform automatic facial landmarking which are sufficient to extract the semantic features. We achieve this by utilizing our proposed AR model (Chapter 3). However, we modified our AR model in this chapter as a component-based model while preserving the amount of landmarks to perform more accurately for extracting better semantic features. In addition, we also involve our proposed robust glasses model (Chapter 6) to detect the presence of glasses on all the faces as one of the semantic features.

In order to obtain the semantic concepts of each face, we prepare some benchmark samples to represent various semantic features of the face (e.g shape of the nose or mouth). All

the faces will be mapped to each of these benchmarks by assigning "membership degree". These memberships are used as features in the face image retrieval phase.

We conducted the experiments based on the **success rate** of finding the correct subject. The result shows that our proposed automatic face images retrieval system can achieve a significant result. We also discovered which semantic features contribute the most and least for face images retrieval.

The structure of this chapter is as follows. Section 7.1 describe more details on the preparation of face images dataset, the improved AR model, which semantic features we use along with the proposed benchmarks and the whole framework of the system. Section 7.2 describes the performance evaluation of our proposed system. Section 7.3 summarizes the contributions made on this chapter.

7.1 Framework

Our semantic-based face images retrieval system consists of three main stages. First, we prepare our face images gallery set from AR database (Martínez and Benavente, 1998). All face images with strong illumination are normalized via Multi-scale Self Quotient image (MSQ) technique (Wang *et al.*, 2004). We then automatically extract all the facial landmarks and glasses presence from all the face images on gallery set via the improved version of our proposed AR model (Chapter 3) and glasses models (Chapter 6). Second, the mapping of facial semantic is done to each face based on the chosen benchmarks. This is done to assign the "membership degree" of each semantic features to each semantic benchmark (e.g narrow eyes, medium eyes, and widely-opened eyes). Lastly, the simulation of semantic query with various scenarios for face images retrieval is conducted. The framework can be seen in Figure 7.1.

7.1.1 Face Database

We chose 117 subjects from AR database (Martínez and Benavente, 1998) as our face image gallery set. Ten images for each subject were selected: two neutrals, two angers, and six illuminated neutrals from both sessions. We still include angry facial expression because our observation revealed that the facial components are not significantly different from neutral expression compared to smile and scream expressions.

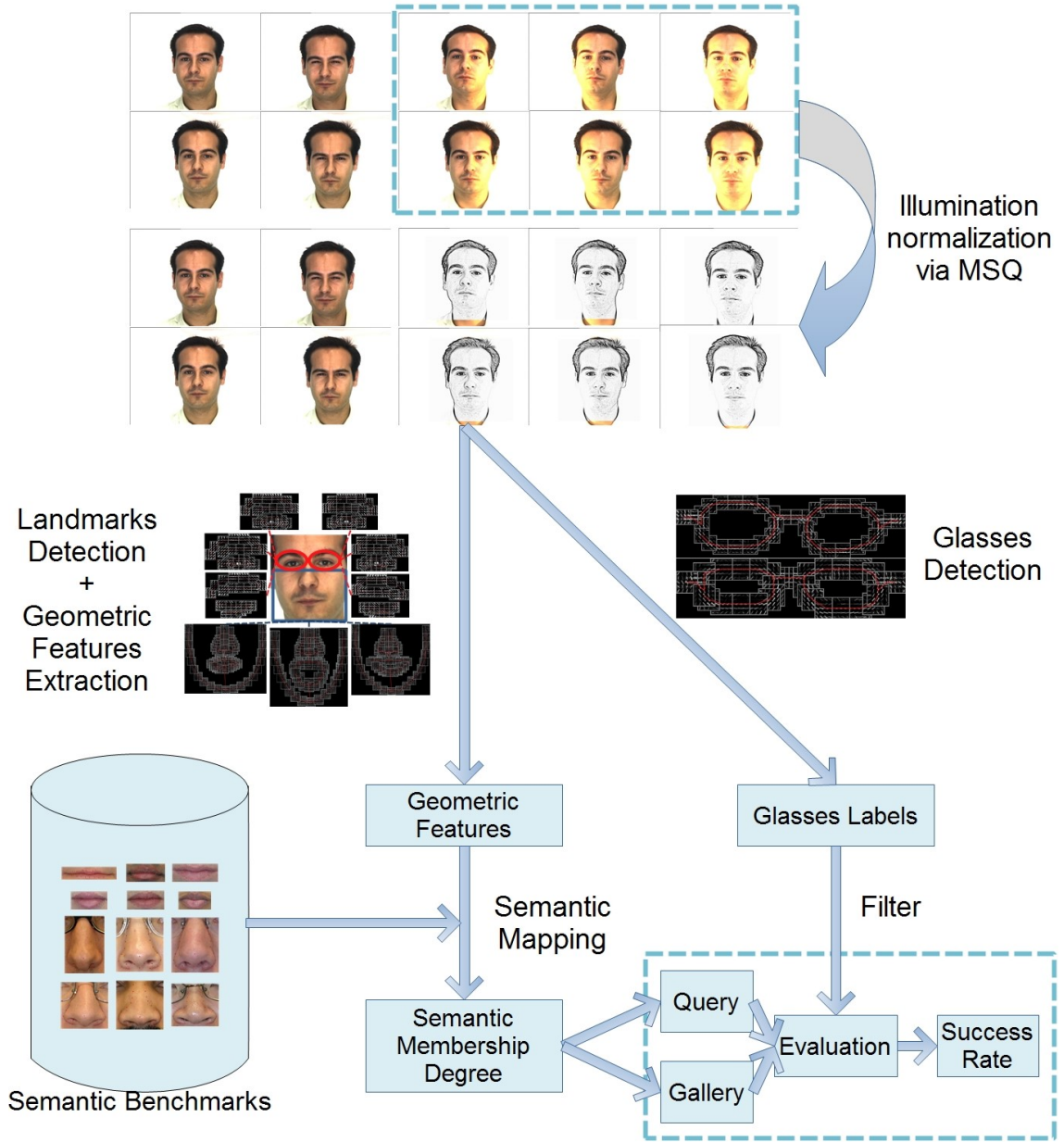


Figure 7.1: The framework of our semantic-based face images retrieval. **(1)** We have 117 subjects with ten images each. Six of them contain various illumination which are normalized through Multi-scale Self Quotient image (MSQ) approach. 130 facial landmarks are then extracted from all the face images through our proposed component-based AR model. Furthermore, glasses presence labels are also created by detecting it through our proposed glasses model. **(2)** Geometric features (e.g eye distance and shape Triangular Area Region (TAR) feature) are extracted based on the facial landmarks information. All these features are mapped semantically to define their "membership degree" to each semantic benchmark sample. **(3)** These membership degree and glasses labels are then used as features to perform face images retrieval.

All six illuminated faces are normalized via the Multi-scale-Self-Quotient-image (MSQ) approach by Wang *et al.* (2004) which is a part of implementation in INface toolbox by Štruc and Pavešić (2009); Štruc and Pavešić (2011). Basically, the Self-Quotient Image approach consists of two main stages: illumination estimation and illumination effect subtraction. Illumination is considered as the extrinsic factor and thus estimated to produce a synthesized image with different albedo mapping. The illumination normalized images are obtained by calculating the difference between the logarithms of original faces and the corresponding synthesized images. The illumination normalization will remove the color information since it is applied on grey-scale images. However, it is not a problem to our proposed AR model since we have shown that the loss of color information does not significantly affect the accuracy as long as the edge information is still clear (Chapter 3.3.4). The examples of the chosen faces and its normalized version are shown in Figure 7.2.

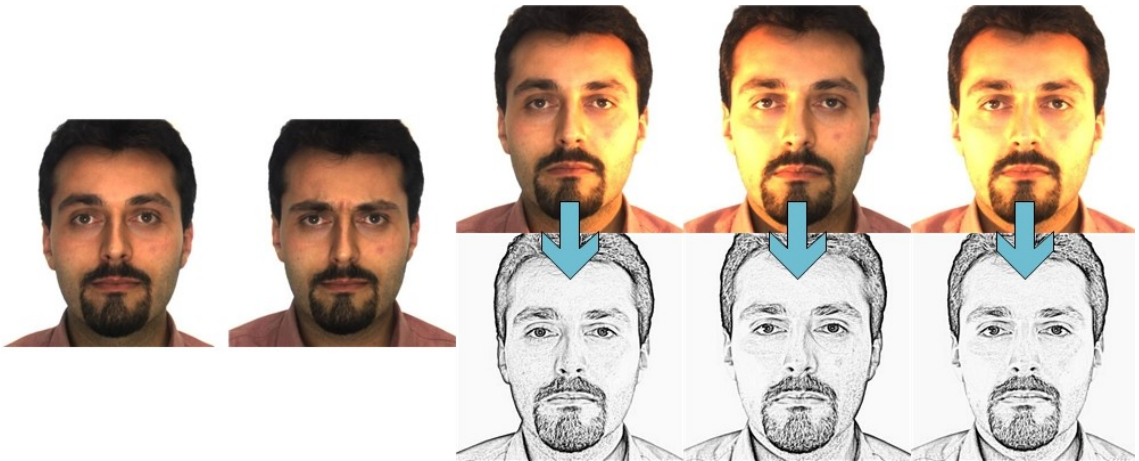


Figure 7.2: Samples of faces from AR database. All illuminated faces are normalized via Multi-scale-Self-Quotient-image (MSQ) approach.

7.1.2 Facial Landmarks Extraction

All facial semantic concepts are extracted through geometric features obtained from facial landmarks. Initially, we planned to retrieve facial landmarks via the previously proposed automatic AR model (Chapter 3). However, despite the significant improvement on the landmarks detection accuracy as shown in Section 3.3, the AR model is still restricted on fixed set of various facial expressions in training (neutral, smile, angry, and scream from AR database). For instance, a face with neutral expression is expected to have fully-opened eyes while smiling faces have slightly-closed/narrow eyes. However, this scenario is not necessarily always true as shown in Figure 7.3 (**Top**). It is possible that people have

narrow eyes even on neutral expression. Similarly, it is also possible that people smile with widely-opened eyes. We believe these restrictions cause a negative impact on the accuracy of retrieved landmarks.

Therefore, we proposed an alternative way of training and using the AR model. We trained the AR model as a **component-based framework** in this chapter. This idea was inspired by the concept of component-based model CompASM by Le *et al.* (2012). We conducted separate model training for left eye, right eye, and lower face regions as shown in Figure 7.3 (**Bottom**). We use the subset of training dataset from AR model (Chapter 3.2) which is the first session of AR database. Only three facial expressions (neutral, smile, and scream(open mouth)) are used to train lower face region. Furthermore, we manually choose 50 face images (as suggested by Zhu and Ramanan (2012a)) for each eye category: widely-open, slightly open/narrow, and closed eyes to train left and right eyes accordingly. The amount of the landmarks are still preserved from the AR model (130 landmarks).

Since our observation revealed that eyes are not necessarily influenced by facial expression, the process of facial landmarks extraction is conducted independently. We first extracted facial landmarks on the lower face region (chin, nose, and mouth). Afterwards, we can focus on the face upper region to localize landmarks on eyes and eyebrows.

We conducted an experiment to evaluate the accuracy improvement of our proposed component-based AR model. We once again employed the standard procedures of evaluating facial landmarks used in 3.3.1 as mentioned by Çeliktutan *et al.* (2013). The *relative error* and *detection rate* on 5%, 10%, and 20% Inter-Ocular Distance (IOD) were measured on 17 fiducial landmarks from the *m17* set. As a reminder, this set refers to eyebrow corners (4 points), eyes corner and centres (6 points), nose tip and both sides (3 points), landmarks around mouth including the corners (4 points). This was conducted on 2 database: AR (second session, neutral and smile) Martínez and Benavente (1998) and PUT Kasinski *et al.* (2008). The summary of the result can be observed in Table 7.1. It has shown that our component-based AR model produces a lower error rate and higher detection even on the lowest IOD.

Table 7.1: Facial landmarking performance improvement with component-based AR model.

	AR database session 2		PUT database	
	AR model	Component-based	AR model	Component-based
Relative Error	0.0362	0.0339	0.0625	0.0600
Detection 05% IOD	79.70 %	83.32 %	52.64 %	56.24 %
Detection 10% IOD	97.03 %	98.00 %	91.96 %	93.49 %
Detection 20% IOD	99.83 %	99.83 %	99.58 %	99.58 %

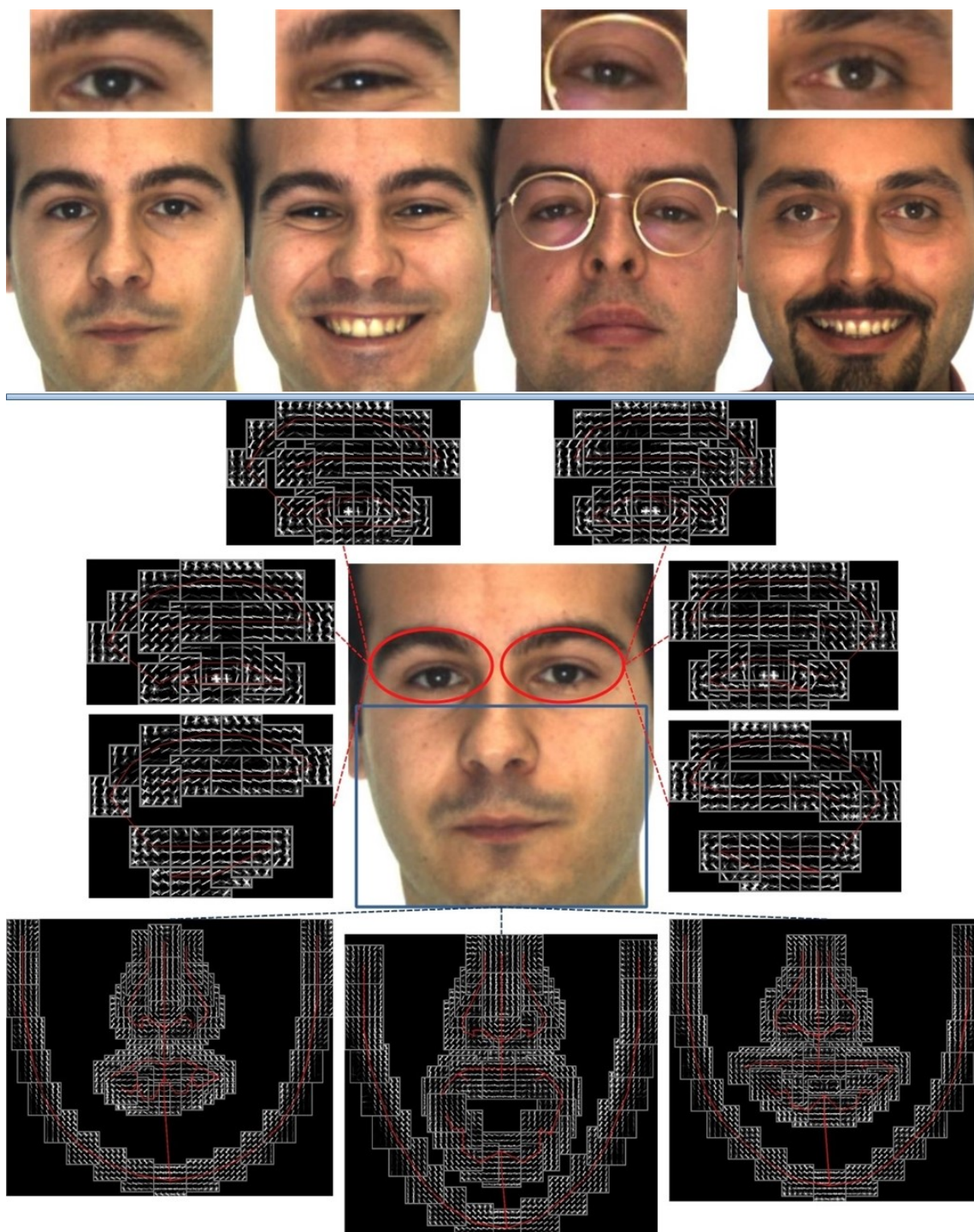


Figure 7.3: **(Top)** Eyes shape/size are not necessarily affected by facial expressions. The eyes might look narrow or wide open on any facial expression. **(Bottom)** Our proposed component-based face models extended from AR model. Landmark fitting for both eyes are not affected by the facial expression on the lower part of the face.

7.1.3 Semantic Features

We selected several semantic features extracted from the detected facial landmarks. We avoid features which are too specific such as curvature of the lower eyelids or shape of the nasal tip (Karczmarek *et al.*, 2015) because these features are too sensitive and relies heavily on the perfect accuracy of the landmarks. Since all the facial landmarks were obtained automatically in our framework, a slight margin of error is to be expected. Instead, we focused on broader description of the facial components (e.g whole nose or mouth).

For each semantic feature, we chose a few image samples as the benchmark for registering membership degree of every face to each semantic category. All the chosen benchmarks were selected from some face images on CMU multiPIE (Gross *et al.*, 2010) and CAS-PEAL-R1 (Gao *et al.*, 2008) database. The idea of having a few benchmark samples is motivated by Ren *et al.* (NA) ¹.

7.1.3.1 Glasses Presence/Existence

The presence of the glasses has been used as one of the semantic features for face retrieval as conducted by Alattab and Kareem (2013) and Vaquero *et al.* (2009). On our face retrieval system, we utilized the proposed tree-structured glasses models from Chapter 6 to distinguish between wearer and non-wearer of glasses automatically (Figure 7.4). The detection rate for all 1170 face images are perfect. It is as consistent as the high performance from Chapter 6.2.1. The assumption here is that the glasses presence is consistent on the same subject (on both query and gallery set). Therefore, this feature can be used to **filter** the gallery set on face images retrieval process for better accuracy.

7.1.3.2 Geometric Ratio

The next features are based on the ratio of the geometric information to describe the distance and size of the eyes (Figure 7.5). By utilizing the ratio, the features are independent to the size of faces, thus makes it easier to compare on any face.

We defined three types of benchmark for eye distance: **close**, **medium**, and **far**. These categories are based on the ratio between inner eye corners distance and the width of the

¹This reference is still under review, so no publication date is available yet

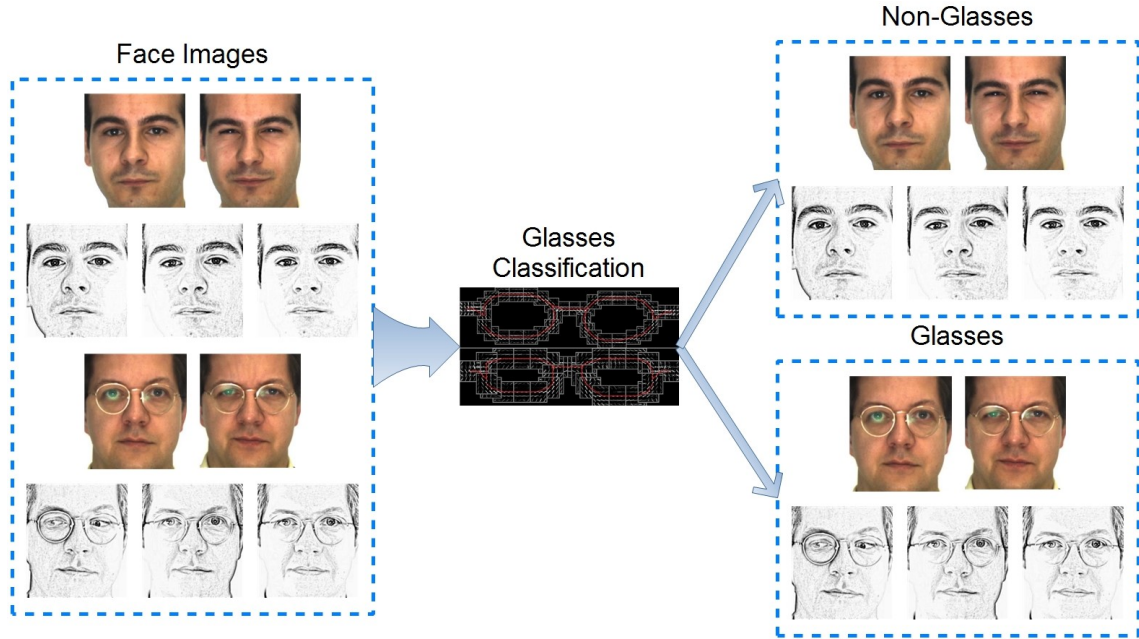


Figure 7.4: All face images are categorized based on the presence of glasses.

eyes (measured as horizontal distance of outer and inner eye corners). It is considered 'close' if the mid-gap is smaller than the corresponding eye width, whereas the opposite case is applied for 'far' category. 'Medium' is only for the face where the mid-gap can fit another eye almost perfectly.

Similarly, we also defined three benchmarks for the size of each eye: **narrow**, **medium**, and **widely-opened**. 12 landmarks were manually marked on each benchmark. The eye size is measured based on the ratio between the *height* and *width* of the eye. The height is calculated as the distance between high and low mid-points of the eye, whereas the width is calculated as the distance between both eye corners.

7.1.3.3 Geometric Shape

The last set of semantic features are based on the 2-dimensional shape description of the facial components. We adopted the same shape feature extraction approach used by Conilione and Wang (2012) to compute Triangular Area Region (TAR) feature (El Rube *et al.*, 2005). TAR feature is considered as an efficient shape descriptor on both computational cost and space/memory requirement. Furthermore, it is also invariant to various factors such as translation, rotation, scale, affine transforms, noise and occlusions (Yang *et al.*, 2008).

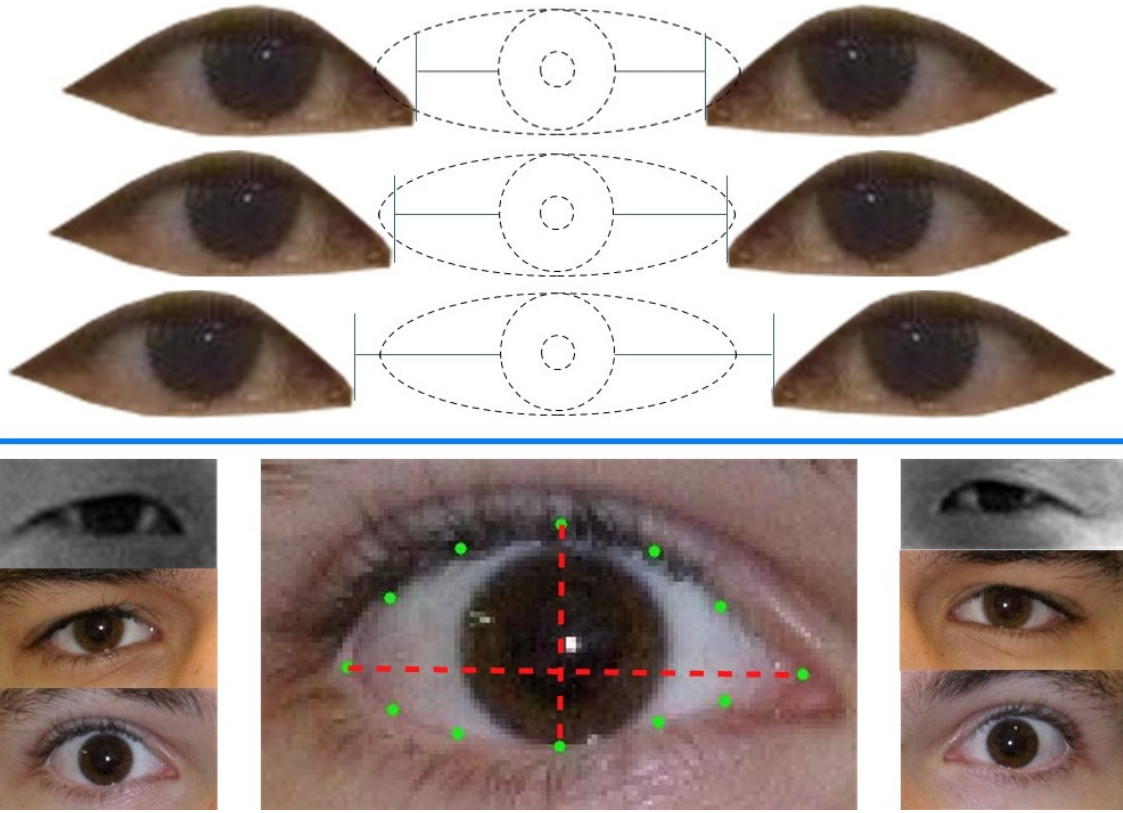


Figure 7.5: **(Top)** The distance between both inner eye corners. We selected three types of benchmarks: close, medium, and far with respect to the width of the eye. **(Bottom)** The size of the eyes calculated through the ratio between its height and width. We have three types of eye size: narrow, medium, and widely-opened for both left and right eyes.

Let $L = (x_i, y_i), i = 1, 2, \dots, N$ be a 2-dimensional object in an image (e.g eyebrow) which is expressed via a set of N points forming a closed contour. Each point (x_i, y_i) is a Cartesian coordinate of i th landmark obtained through the facial landmark detector. The shape of the object are described as the collection of Triangular Area Region (TAR) between 3 equal-distant landmarks $(x_{i-t}, y_{i-t}), (x_i, y_i), (x_{i+t}, y_{i+t})$ on all N landmarks along the contour where t is the length of the triangle (e.g $t = 1$ means 3 neighboring landmarks). The formula for the triangle area is defined as:

$$TAR(i, t) = \frac{1}{2} \begin{vmatrix} x_{i-t} & y_{i-t} & 1 \\ x_i & y_i & 1 \\ x_{i+t} & y_{i+t} & 1 \end{vmatrix}$$

where the sign of the TAR depends on:

$$TAR(i, t) = \begin{cases} = 0 & \text{if straight line} \\ < 0 & \text{if convex contours} \\ > 0 & \text{if concave contours} \end{cases}$$

Straight
Line

Convex
Contour

Concave
Contour

The value of t ranges from 1 to $\lfloor (N-1)/2 \rfloor$ due to the constraint by the periodicity of the closed loop of L . The boundary condition $t = N/2$ is defined as:

$$TAR(i, t) = \begin{cases} 0 & \text{if } t = \frac{N}{2}, N \text{ is even} \\ \text{undefined} & \text{if } t = \frac{N}{2}, N \text{ is odd} \end{cases}$$

We can regard the value of TAR for any t as an individual scale space function (Yang *et al.*, 2008). Therefore, by combining all the TAR value for $t = [1, \dots, (N-1)/2]$, we define a multi-scale space TAR feature to describe the shape of an object. In this experiment, we selected a few facial components shape benchmarks. Each benchmark sample is manually landmarked accordingly to extract its TAR feature. This is the only part of our system where manual landmarking is still needed because these benchmarks act as the ground truth for comparison. However the scale is much smaller compared to having to manually create landmarks on all face images which can be prohibitively time consuming. All the facial component shape benchmarks (with the corresponding TAR) we choose are as follows:

- 3 chin shapes (21 landmarks) (Figure 7.6).
- 4 eye (left and right) shapes (12 landmarks each) (Figure 7.7).
- 5 eyebrow (left and right) shapes (12 landmarks each) (Figure 7.8).
- 6 mouth shapes (20 landmarks) (Figure 7.9).
- 6 nose shapes (27 landmarks) (Figure 7.10).

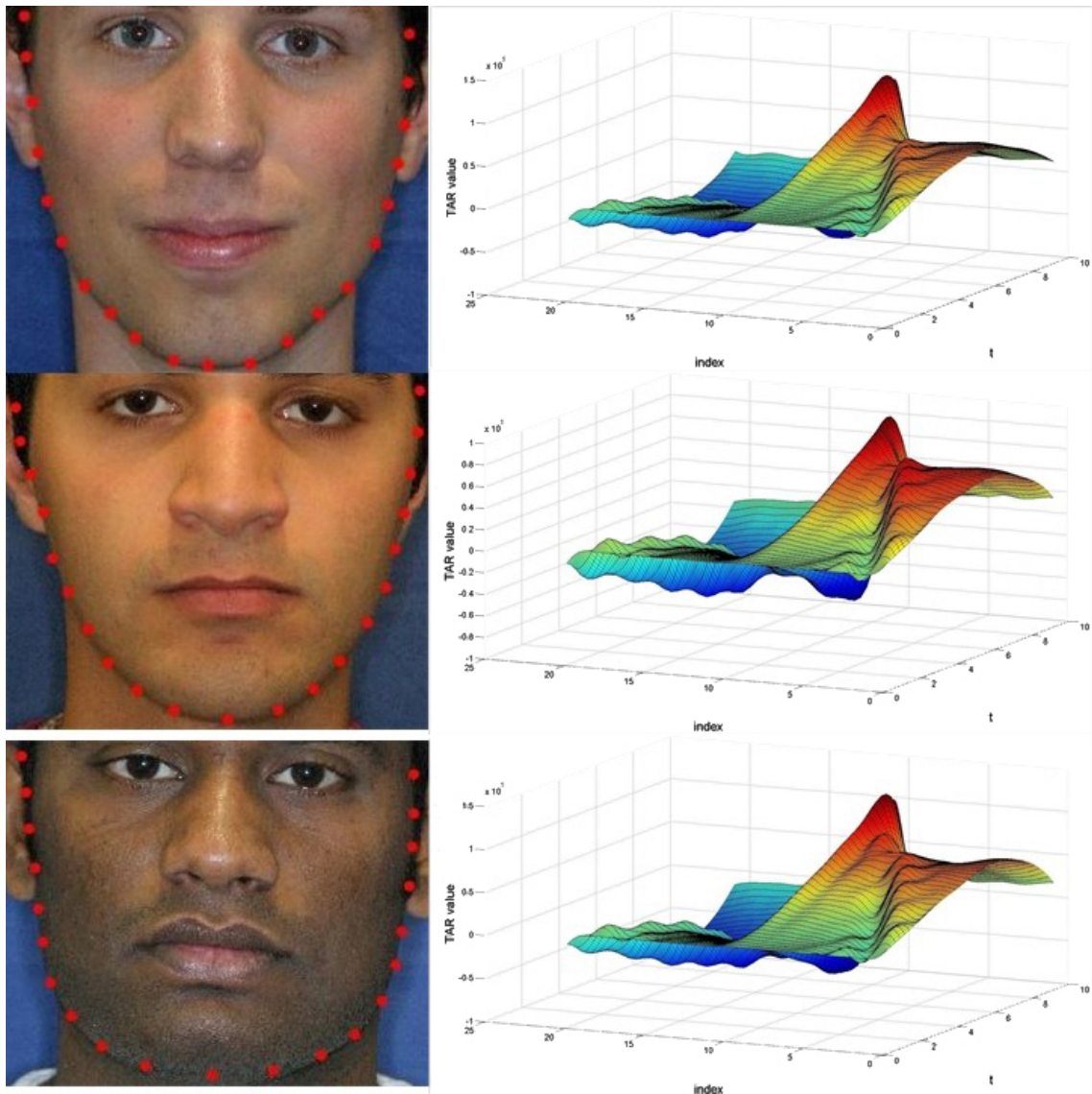


Figure 7.6: Three chin shapes.

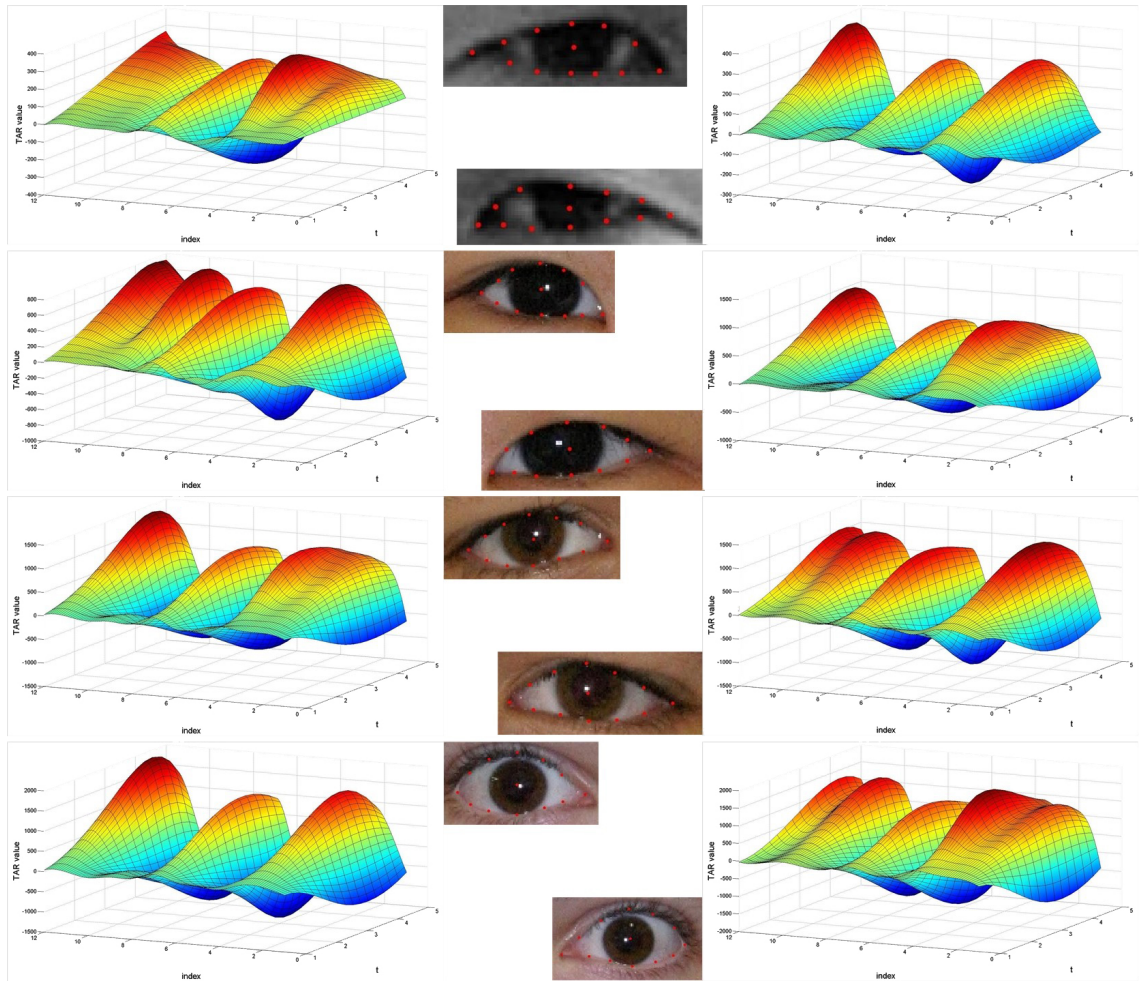


Figure 7.7: Four right and left eye shapes.

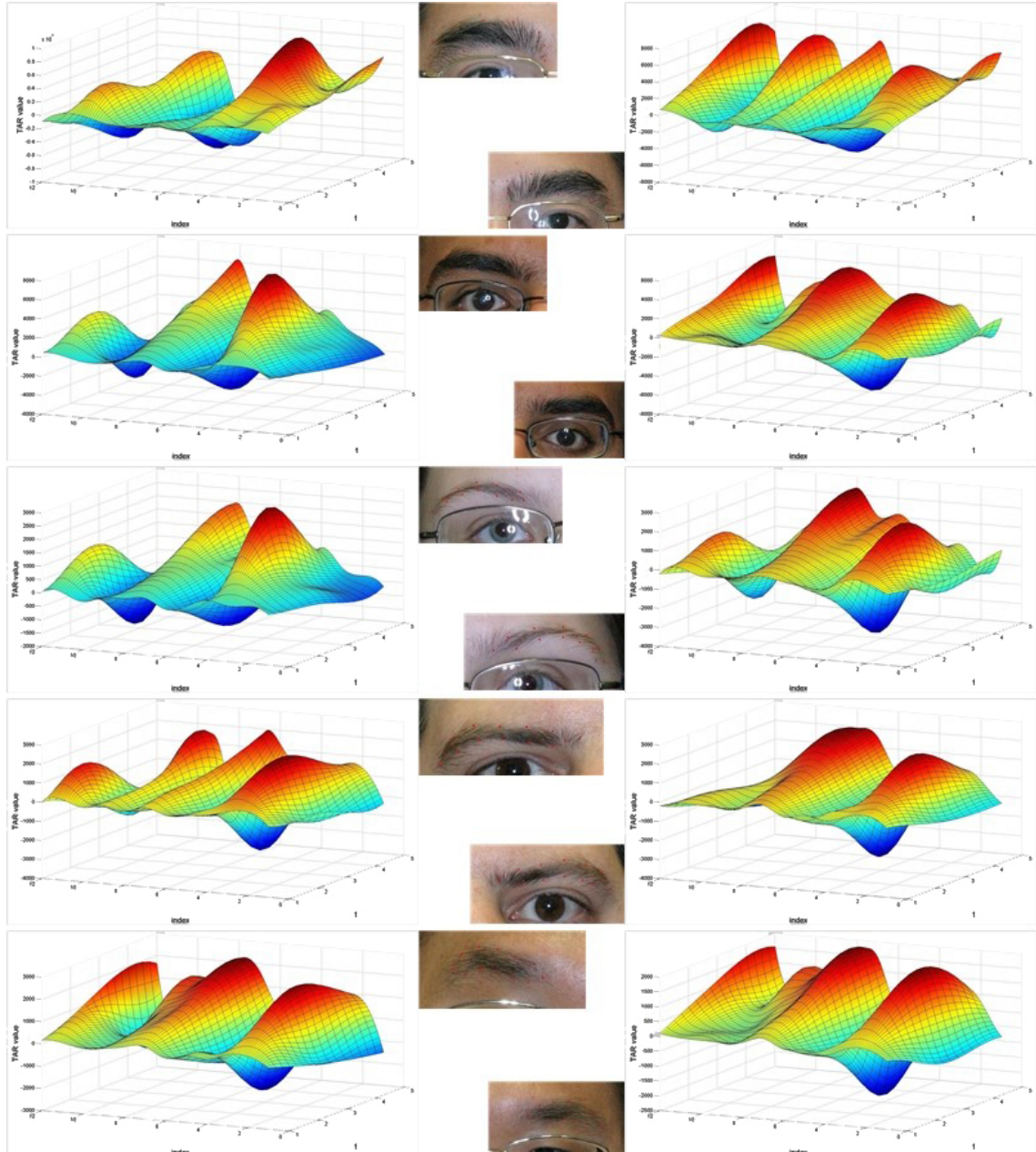


Figure 7.8: Five right and left eyebrow shapes.

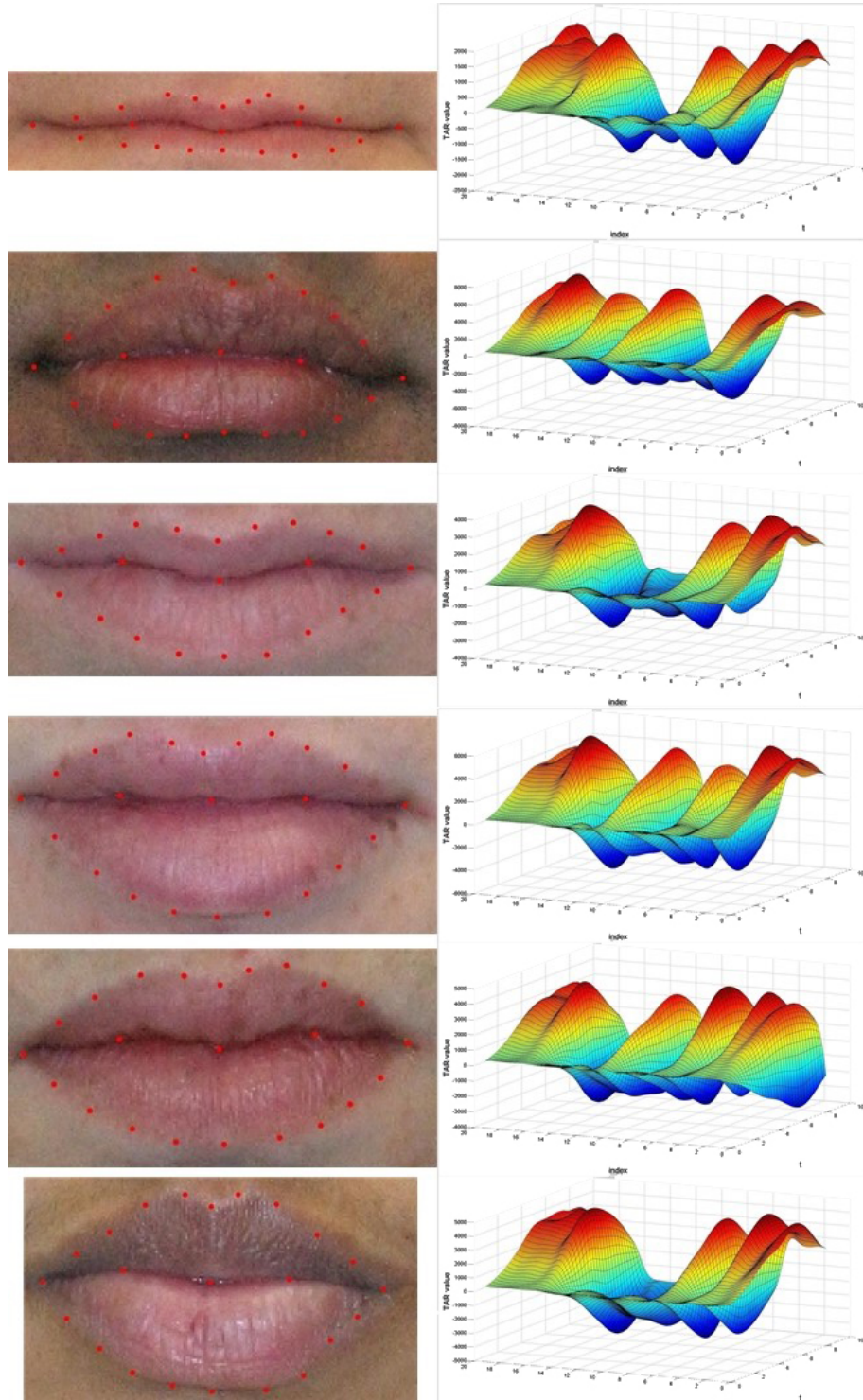


Figure 7.9: Six mouth shapes.

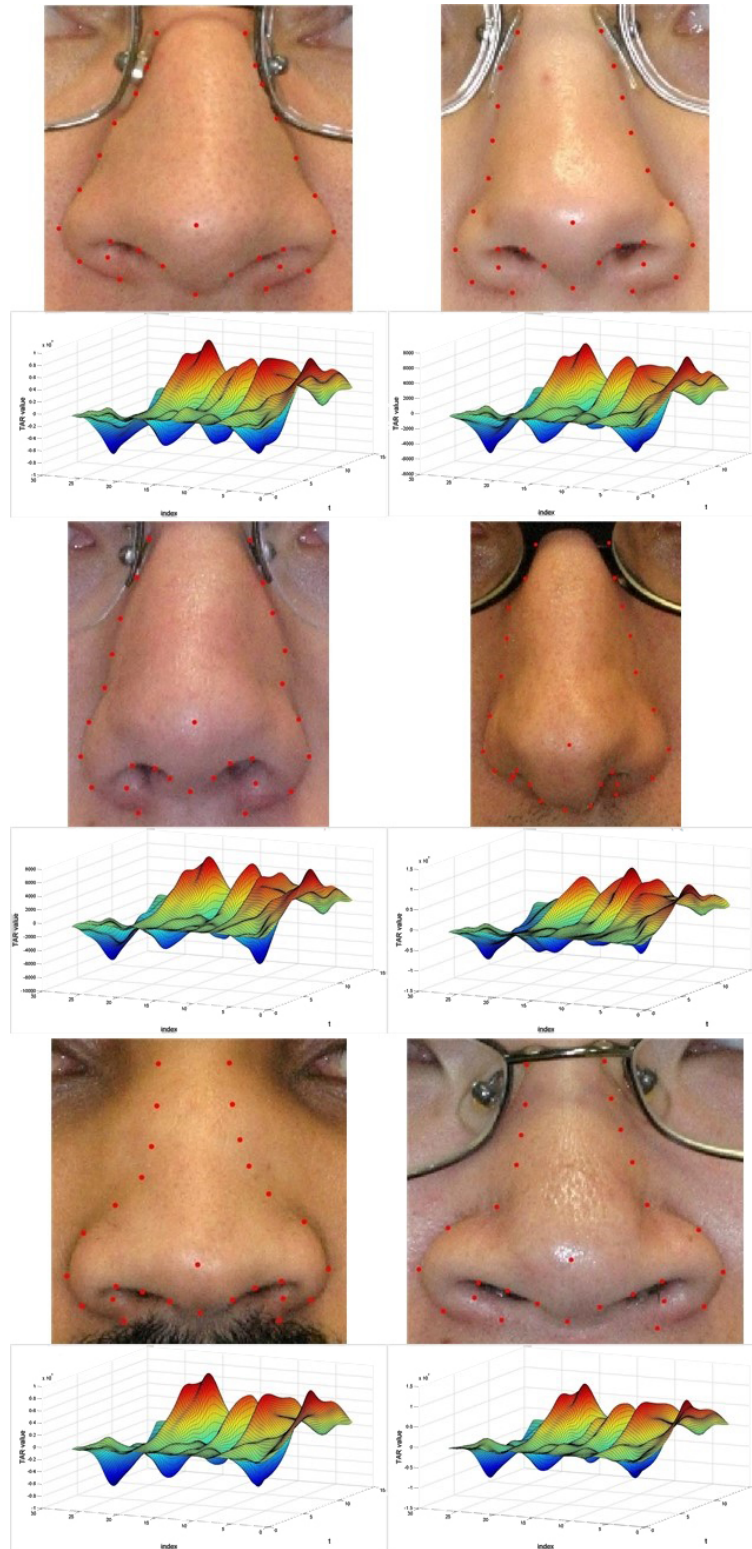


Figure 7.10: Six nose shapes.

7.1.4 Semantic Mapping

After we extracted facial landmarks out of all 1170 face images (117 people x 10 images each), we compute all the geometric features as mentioned previously and compare them systematically to calculate their "membership" degree for each feature benchmark. The magnitude of the membership degree depends on how close/similar the feature to each benchmark. We consider two types of comparisons:

- Distance: This applies for eyes distance and size of eyes. The similarity is calculated based on the absolute distance. For each semantic category (e.g right eye size), the magnitudes are normalized into $[0, 1]$ where 1 defines perfect resemblance (zero distance) and 0 defines the furthest distance among all 1170 faces.
- Correlation: This applies for all multi-scales space TAR features on chin, eye, eyebrow, mouth, and nose. The similarity is calculated based on the Pearson correlation coefficient between TAR features (Fisher, 1958; Kendall and Stuart, 1979; Press *et al.*, 1992). Let A and B be the TAR features of a facial component from two face images, then the correlation coefficient $P(A, B)$ is calculated as:

$$P(A, B) = \frac{1}{N-1} \sum_{i=1}^N \left(\frac{A_i - \mu_A}{\sigma_A} \right) \left(\frac{B_i - \mu_B}{\sigma_B} \right)$$

where N is the dimension of TAR A or B (same length). μ_A and μ_B are the mean of A and B respectively. σ_A and σ_B are the standard deviation of A and B respectively. This formula can be simply written in terms of covariance of A and B as:

$$P(A, B) = \frac{cov(A, B)}{\sigma_A \sigma_B}$$

For each semantic category (e.g chin shape), the magnitudes are normalized into $[0, 1]$ where 1 defines perfect similarity (correlation = 1.0) and 0 defines the worst correlation (usually a negative number) among all 1170 faces.

In order to gain a better understanding intuitively, let us see some examples. For instance, the benchmark for right eye size are *Reye_size_benchmark* = $[0.2607, 0.3697, 0.4866]$ for narrow, medium, and widely-opened. As a reminder, these values are calculated as the ratio between the height and width of the right eye benchmark samples. Assuming we

only have 3 face images with the right eye size value $gallery_Reye_size = [0.28, 0.38, 0.45]$, thus we calculate the distance matrix D as follows:

$$D = \begin{bmatrix} |0.2607 - 0.28| & |0.3697 - 0.28| & |0.4866 - 0.28| \\ |0.2607 - 0.38| & |0.3697 - 0.38| & |0.4866 - 0.38| \\ |0.2607 - 0.45| & |0.3697 - 0.45| & |0.4866 - 0.45| \end{bmatrix}$$

$$D = \begin{bmatrix} 0.0193 & 0.0897 & 0.2066 \\ 0.1193 & 0.0103 & 0.1066 \\ 0.1893 & 0.0803 & 0.0366 \end{bmatrix}$$

on which D will be normalized $D = \frac{D}{\max(\max(D))}$ as such:

$$D = \begin{bmatrix} 0.0934 & 0.4342 & 1.0000 \\ 0.5774 & 0.0499 & 0.5160 \\ 0.9163 & 0.3887 & 0.1772 \end{bmatrix}$$

However, since low distance means strong similarity, we reverse the value of D by subtracting the value of 1 out of it as $D = 1 - D$:

$$D = \begin{bmatrix} \mathbf{0.9066} & 0.5658 & 0 \\ 0.4226 & \mathbf{0.9501} & 0.4840 \\ 0.0837 & 0.6113 & \mathbf{0.8228} \end{bmatrix}$$

where each row represents the membership degree of each face to the three benchmark (**narrow, medium, and widely-opened**) of the right eye size. The same concept can be applied on TAR features. However, the distances are based on the correlation. Involving all the membership degrees from all semantic features produces a vector of semantic features for each face of size 42 ($42 = 3$ eye distance + 3 chin shapes + 10 eyebrow shapes (left and right) + 6 eye sizes (left and right) + 8 eye shapes (left and right) + 6 mouth shapes + 6 nose shapes).

These membership degrees can be used to describe the *semantic concepts* of each face. In this example, each face has a dominant value which represents strong similarity to one of the benchmarks (e.g first face has narrow right eye, second face has medium right eye,

and so on...). However, it is also possible for a face not having any significantly dominant value. For instance, if the size of the eye is perfectly in the middle between narrow and medium size, then it can be described with a fuzzy manner such as "rather narrow".

The advantage of this semantic mapping approach is that it is computationally efficient and allows for easy semantic benchmarks expansion in the future since it does not require large amount of samples per benchmark. For instance, we can easily add more types of eye shapes with just one image sample each. Furthermore, it is also possible to add a completely new semantic feature such as forehead size or shape.

7.2 Experiments

7.2.1 Experiment Setup and Performance Evaluation

As the main purpose of face images retrieval is to find the face(s) with the same identity as the query image, we evaluate the performance of our face retrieval system based on the **success rate** of finding the correct identity among the top k results. In our experiment, we choose $k = 5$. We can not directly compare our approach to others numerically due to various factors such as their facial landmarking by hand (manual) and difference on semantic features.

We divided the face images from 117 subjects into two sets. The first set containing 50 subjects are evaluated in order to learn the best combination of semantic features based on its success rate. We can observe which semantic features contribute more for semantic-based face images retrieval. The learned combinations are then used to evaluate the performance of the remaining 67 subjects. The semantic feature combinations are learned through the greedy approach as used by Li *et al.* (2011). The basic idea is that we initially evaluate the performance with a single semantic features (e.g mouth shape) iteratively and record the average success rate. We then choose the one with the best result and proceed to find the next best combination with two semantic features and so on. This process is repeated until all semantic features are involved. The combination with the highest success rate is chosen to perform face images retrieval on the second set for the remaining 67 subjects.

Since each subject consists of 10 face images, we choose randomized n images as the simulated semantic queries and the remaining $(10 - n)$ images as the gallery set. For our experiment, we choose $n = 1, 2, 3, 4, 5$. It means that we have 5 experiments with 1, 2, 3,

4 and 5 queries. We iteratively evaluate the performance 100 times for each n and each semantic combination learning to obtain the average success rate. The classification is conducted via the subspace projection technique LDA (Belhumeur *et al.*, 1997) to learn the projection matrix W from the gallery set. Furthermore, with the information of **glasses presence labels**, we ensure the system to compare only with the subject with the same state to **filter** some gallery images. Afterwards, The retrieval result shows the top 5 closest distance to the query.

7.2.2 Experiment Results

We evaluate this experiment in two scenarios:

- For each subject, we compute the average of his/her gallery images (excluding query) as the new semantic representation of that subject. It means that each person will have only one representation in the gallery for calculating transformation matrix W with LDA and perform top 5 face images retrieval with the query.
- For each subject, we also compute the average of his/her gallery images (excluding query) as the new semantic representation of that subject. However, we still keep the original gallery images along with its average to compute transformation matrix W and perform top 5 face images retrieval. In this case, every subject will have an additional representation in the gallery set. We make this setup more challenging by providing more selection in the gallery while still retrieving only top 5 face images.

All the semantic features can be divided into ten categories. We assign each category with a single number from 1 to 10 for easy identification as follows:

(01) Eye distance	(02) Chin shape
(03) Left eyebrow shape	(04) Left eye size
(05) Left eye shape	(06) Mouth shape
(07) Nose shape	(08) Right eyebrow shape
(09) Right eye size	(10) Right eye shape

7.2.2.1 Scenario 1

We first learn the optimal semantic combination of the first 50 subjects with various number of queries. The result is summarized in Table 7.2 for n queries. As a reminder, this means we choose randomized n face images per subject as query set and the remaining $(10 - n)$ face images per subject as gallery set. In this experiment, we show the result when $n = 1, 2, 3, 4, 5$. This table shows the success rate of the face images retrieval starting from using only one semantic category up to ten categories. The order of the chosen semantic categories depends on the highest success rate achieved for each combination. For instance, Table 7.2 shows that nose shape (07) alone can achieve 48.90% success rate for 1 query per subject. Afterwards, combining the nose shape (07) with eye distance (01) improves the result into 62.54%. This step is repeated until all semantic categories are involved. Overall result can achieve the highest success rate close to 80%.

Table 7.2: Result of learning semantic combination on the first 50 subjects based on the success rate. These are evaluated with $n = 1, 2, 3, 4, 5$ queries.

Semantic Amount	1 Query		2 Queries		3 Queries		4 Queries		5 Queries	
	Rate	Sem.	Rate	Sem.	Rate	Sem	Rate	Sem.	Rate	Sem.
1	48.90	07	48.08	07	48.65	07	47.75	07	46.97	07
2	62.54	01	62.61	01	61.91	01	61.35	01	60.76	01
3	71.48	06	70.64	06	69.96	06	68.89	06	67.44	06
4	77.62	08	75.96	08	74.74	08	73.12	08	71.76	08
5	76.96	04	77.06	09	76.46	09	74.53	09	72.70	09
6	79.64	10	77.95	10	76.87	10	75.44	04	74.06	04
7	79.66	09	78.69	04	77.76	04	76.75	10	75.01	03
8	78.52	03	79.13	03	77.66	03	76.78	03	75.29	10
9	78.82	05	78.26	05	76.74	05	76.04	05	74.33	05
10	74.26	02	73.68	02	72.63	02	71.23	02	69.74	02

Table 7.3: The retrieval success rate on the remaining 67 subjects.

Chosen Semantics	1 query	2 queries	3 queries	4 queries	5 queries
All except Chin Shape and Left Eye Shape	82.09	79.10	78.61	75.75	74.33

The result shows that not all semantic categories bring positive impact on face retrieval result. All five cases ($n = 1, 2, 3, 4, 5$ queries) show that the involvement of shape of chin and left eye shape (02 and 05) decrease the success rate. It is especially bad when chin shape is involved. There is a possibility that the shape information on chin are not sufficiently discriminative since they are quite similar to each other. Left eyebrow shape (03) also occasionally decrease the performance, however it has much less impact. On the other hand, the shape of nose, mouth and right eyebrow along with the distance between eyes (07, 06, 08 and 01) are the features with the most contribution towards success rate.

This result is consistent with the discovery by Conilione and Wang (2012) which states that nose information contributes the most.

However, it seems peculiar that the left eyebrow shape (03) does not contribute as much as right eyebrow shape (08). It is possible that the contribution of left eyebrow has been overshadowed by the right eyebrow. Since it is high likely both eyebrows have similar shape (although reversed horizontally), the semantic information of one of them is already sufficient. Therefore, adding a similar feature will contribute less new information. A similar pattern can also be observed from the shape of right eye (10) and left eye (05). Once one of them (right eye) contributes to the retrieval result, the other eye (left eye) contributes less. In this example, left eye shape even decreases the performance.

Based on the learnt combination from the previous 50 subjects, we perform another face images retrieval on the remaining 67 subjects without involving the shape of chin and left eye (02 and 05). Table 7.3 shows that the success rate can achieve significant result up to 82.09% success rate from the learned semantic combination. It can be seen that the result is gradually decreasing as the amount of queries increases. This can be justified by the fact that the amount of gallery images becomes less as the query increases per subject. This implies that we get less and less information of each subject in the gallery set while we get more variation of queries to be tested. However, even with 5 queries and 5 galleries per subject, we still can achieve 74.33%.

The experiment results in Table 7.2 and 7.3 involves glasses filter when conducting face images retrieval. We want to observe how the glasses presence filter help improving the success rate. Table 7.4 shows the comparison between the non-involvement and involvement of glasses filter. It can be seen the glasses filter significantly improves the success rate approximately **10% to 13%**.

Table 7.4: Success rate improvement before and after glasses filter.

Success Rate on the First 50 Subjects					
	1 query	2 queries	3 queries	4 queries	5 queries
No Glasses Filter	69.34	67.45	66.59	65.35	63.50
With Glasses Filter	79.66	79.13	77.76	76.78	75.29
Success Rate on the Remaining 67 Subjects					
	1 query	2 queries	3 queries	4 queries	5 queries
No Glasses Filter	71.64	67.91	66.17	63.43	61.49
With Glasses Filter	82.09	79.10	78.61	75.75	74.33

7.2.2.2 Scenario 2

The result for semantic combination learning is summarized in Table 7.5. Once again, this table follows the same format as previous scenario. We still can see similar pattern such as the shape of nose, mouth and right eyebrow are some of the biggest contributors toward the success rate. Furthermore, similar pattern of "overshadowing phenomena" between pairs of eyebrow shapes, eye shapes, and eye sizes still presents. For example, on the case of 1 query, when left eyebrow shape (03) is involved beforehand, the right eyebrow shape (08) contributes less improvement (from 6.32% to 4.40%). Similarly on the case of 2 queries, the improvement by the right eyebrow (7.94%) has been reduced after involving left eyebrow afterwards (drop to 3.54%). However, the difference is that involving all semantic features does not significantly impair the retrieval result like in the previous scenario.

Based on the result of Table 7.5, we decided to include all semantic features for face retrieval on the remaining 67 subjects. Once again, the result in Table 7.6 shows significant result by achieving 80.60% success rate for the highest result by involving all semantic features.

Table 7.5: Result of learning semantic combination on the first 50 subjects based on the success rate. These are evaluated with $n = 1, 2, 3, 4, 5$ queries.

Semantic Amount	1 Query		2 Queries		3 Queries		4 Queries		5 Queries	
	Rate	Sem.	Rate	Sem.	Rate	Sem	Rate	Sem.	Rate	Sem.
1	52.70	07	50.23	07	49.35	07	47.45	07	46.16	07
2	56.44	06	55.75	01	53.79	01	52.87	01	51.22	01
3	64.02	03	62.72	06	62.49	06	61.53	06	59.41	06
4	70.34	08	70.66	08	69.53	08	67.55	08	65.91	08
5	74.74	01	74.20	03	72.81	03	71.21	03	69.39	03
6	77.22	09	76.10	09	74.87	09	72.74	09	70.65	09
7	79.36	02	77.39	10	75.70	02	73.94	02	71.79	10
8	80.26	10	78.05	02	76.47	10	75.60	10	73.18	04
9	79.86	05	78.27	05	77.11	04	75.56	04	73.30	02
10	79.88	04	78.73	04	77.84	05	75.45	05	73.40	05

Table 7.6: The retrieval success rate on the remaining 67 subjects.

Chosen Semantics	1 query	2 queries	3 queries	4 queries	5 queries
All	80.60	79.85	77.61	76.12	73.43

7.3 Summary

In this chapter, we proposed an automatic face images retrieval based on the retrieved facial landmarks from our proposed component-based AR model and previously proposed glasses models. The whole framework begins from automatic facial landmarks extraction (and glasses detection) followed by automatic semantic mapping to each face and concluded with the query simulation. We begin with our first contribution on proposing component-based AR model. This model has improvement in terms of landmarks accuracy and detection rate. Furthermore, it is less affected by facial expressions. With this component-based AR model, we can automatically extract geometric features from the landmarks which will be mapped as semantic features. Our second contribution is the proposed semantic mapping system and the benchmarks samples. This system can efficiently assign semantic "membership degree" of each geometric feature to each corresponding benchmark samples. Furthermore, our proposed semantic mapping system allows for easy expansion of new samples or completely new semantic features in the future. The third contribution is the usage of glasses presence label detected with our previously proposed glasses model. We utilize the information of glasses presence to filter the result of face images retrieval. We assume that each subject always wears glasses (OR not wearing glasses) all the time on both query and gallery set. This filter will eliminate some choices in the gallery set which leads to higher success rate on finding the query subject.

Our experiment results reveal that our automatically-gained semantic features can be used to achieve significantly high success rate on face retrieval. Furthermore, we also learn that the eye distance and shape of nose, mouth, and eyebrows contributes the most on the result. On the other hand, chin shape contributes the least due to its slight invariance.

Chapter 8

Conclusions and Future Directions

This thesis addresses the problem of improving the performances of automatic frontal faces landmarking system with the application on semantic-based face images retrieval. All the proposed approaches are the further developments of the pictorial-tree-structure face models by Zhu and Ramanan (2012a) described in Chapter 2. Our main contributions reside in the context of accuracy, resolution range, and efficiency via preceding face detection. In addition, an alternative usage of the model was proposed for robust glasses detection/landmarking which can be used to define another facial semantic feature. Lastly, we integrate both facial and glasses landmarks detector to propose an efficient automatic semantic-based face images retrieval framework.

We begin our research with a contribution via developing a face model with higher *accuracy* and *amount of landmarks* in Chapter 3. We achieved this by employing a new facial structure with a high density of facial landmarks inspired by Milborrow and Nicolls (2008). This notion leads to a higher accuracy due to a better landmarks fitting, thus potentially providing better semantic facial features. We refer this proposed model as the **AR model** since it is trained on frontal faces (four different expressions) from AR database. AR model contains close to double amount of landmarks compared to face models proposed by Zhu and Ramanan (2012a). We conduct performance evaluations with a few state-of-the-art approaches based on the relative error and detection rate of the landmarks and the accuracy of the geometric descriptions derived from them. The experiment results reveal a significant overall improvement by our proposed AR model. Lastly, we investigated the effects of various colour spaces on AR model. Due to the slight accuracy change, we concluded that there is no major impact from the colour information as long as the edge information is clear.

We then develop the proposed face models further to cover various face resolutions in Chapter 4. As AR model is trained on large faces, it can only fit the landmarks well on high resolution faces. We proposed to extend AR model via training the face models on **Multi Resolutions (MR)** models to cover low resolution faces. We decided to train MR models on other four scales: 210x210, 150x150, 90x90, and 30x30. As the initial landmarks are too dense for low resolution faces, we designed an automatic adaptive landmarking

framework to preserve important landmarks depending on the size of training faces. We evaluated the performance of MR models on PUT database. We first compare it with Share-146 model by Zhu and Ramanan. Our MR models outperform Share-146 by a significant margin and are able to detect faces as small as 30x30 on which Share-146 would fail. We then compare with two other state-of-the-art approaches: Intraface (Xiong and De la Torre, 2013) and STASM (Milborrow and Nicolls, 2014). The experiment results reveal that MR is comparable on large faces, but slightly less accurate on small faces. However, additional experiment shows that our proposed MR models are more robust and stable against landmarks misalignment in the presence of hair and beard. Furthermore, MR models are less sensitive to false face detection since it can detect the face itself.

We then divert our attention to face images taken in uncontrolled environment in Chapter 5. We propose a novel face detection model called the **Tree-structured Filter Model (TFM)**. The main purpose of TFM is to filter false face detections from the Viola Jones face detector (Viola and Jones, 2004) while preserving high rate of correct detections. TFM is trained on low resolution faces with restricted landmarks and expressions just sufficient to depict intuitive description of frontal human faces, thus making it highly efficient. We also design a complete facial landmarking system by integrating Viola Jones face detector, TFM, and MR models for images taken in uncontrolled system. The experiments are conducted on two uncontrolled databases with the focus on frontal/near-frontal faces. The first experiment demonstrates a significant performance of TFM on maintaining high correct face detections with the lowest false detections. The second experiment shows the advantages of our proposed facial landmarking system compared to other algorithms in terms of detection rate and processing time.

As glasses can be considered as a part of a human face, we extend our landmarking technique into detecting glasses landmarks in Chapter 6. We have two main contributions made in this chapter. The first contribution is the proposed robust **glasses model** which is able to detect and extract 39 glasses landmarks. This tree-structured model is trained from 100 manually selected glasses images from CMU multiPIE database (Gross *et al.*, 2010). We systematically provide 39 glasses landmarks for each training image to ensure high consistency and accuracy. This model is tested on various databases and proven to be remarkably robust on detecting glasses presence along with its landmarks. The second contribution is the proposed automatic integrated glasses removal system to improve face classification performance. We employ two image reconstruction approaches NLCTV inpainting (Duan *et al.*, 2015) and SFDAE Deep Learning model (Pathirage *et al.*, 2015) as a hierarchical double-layered filter to remove the presence of glasses based on the location information extracted by our proposed glasses model. The experiment results demonstrate the robust improvement on both face recognition and verification.

To conclude the thesis, we design an **automatic semantic-based face images retrieval system** based on the landmarks extracted from our proposed component-based AR model and previously proposed glasses model in Chapter 7. Our first contribution is the component-based AR model. We divided AR model into 3 tree-structured models to reduce the effect of facial expressions on eye regions. The experiment results demonstrate improvement in terms of landmarks accuracy and detection rate compared to the original AR model. Our second contribution is the proposed semantic mapping system and benchmark samples. Semantic "membership degree" can be efficiently mapped for each geometric feature to the corresponding benchmark samples. This system also allows for easy expansion of benchmark samples by providing additional samples or even entirely new semantic features in the future. Our last contribution is the utilization of glasses presence information detected by our glasses model to filter the result of face images retrieval. With the assumption that any face subject always wears glasses (or vice versa) all the time on both query and gallery set, the filter is able to eliminate some negative options in the gallery set. The experiments show the huge advantage of utilizing glasses filter on improving the success rate. Lastly, the results also prove that the semantic features extracted automatically from our proposed component-based AR model can be used to achieve significant success rate in face images retrieval.

8.1 Future Study

Despite all the significant performances achieved by all of our proposed approaches, we still have the following possible problems to solve in near future.

- Our face landmarking models are still restricted to frontal faces due to our focus on semantic-based face images retrieval. We believe the concept of high-density face models, adaptive landmarks, and light-weight face filters could be applied to train faces with various angles/poses.
- We believe that the components-based AR model proposed in Chapter 7 is more efficient and robust for facial landmarking, but it deserves further investigations on various types of tree structure.
- For the case of multi-resolutions facial landmarking, we can consider to involve image enhancement techniques to improve the detail of facial features (e.g edge information) for better facial landmarks detection.

- We can expand our proposed semantic-based face images retrieval to involve more facial databases. Furthermore, we also can consider the uncontrolled environment scenario by integrating it with our proposed MR models and TFM. Lastly, further investigation is needed to explore more complex types of facial semantic features such as skin color information.
- Lastly, our proposed approaches are still far from real-time system due to large processing involved in landmarks fitting. Even though we have improved the efficiency through our proposed TFM and fixed-size scaling, there are still rooms for improvement. For instance, we can reduce the features domain by restricting the features pyramid just on the similar scale levels of the original image after the faces are detected by Viola Jones detector and TFM. Therefore, we can avoid excessive computation since early face detection informs us on approximate size of the face. Furthermore, the concept of part sharing (Torralba *et al.*, 2007) adopted in Zhu and Ramanan’s face models can also be applied on our proposed approach. Time complexity analysis conducted by Zhu and Ramanan (2012a) reveals that the computational cost of pictorial-tree-structured model is affected by four factors: Amount of landmarks on each model L , amount of trained models M , feature dimension D and candidate part locations N which bring to performing complete landmarks fitting on the whole image as $O(DNML)$. By conducting comprehensive part sharing, we are able to decrease the number of unique landmark templates which effectively reduce M significantly for computing efficiency.

Bibliography

- (2012). *MATLAB and Statistics Toolbox Release 2012b*. The MathWorks Inc., Natick, Massachusetts, United States.
- Akagunduz, E. and Ulusoy, I. (2007). 3d object representation using transform and scale invariant 3d features. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.
- Akakin, H. C. and Sankur, B. (2007). Automatic and robust 2d/3d human face landmarking. In *2007 IEEE 15th Signal Processing and Communications Applications*, pages 1–4. IEEE.
- Alattab, A. A. and Kareem, S. A. (2013). Semantic features selection and representation for facial image retrieval system. In *Intelligent Systems Modelling & Simulation (ISMS), 2013 4th International Conference on*, pages 299–304. IEEE.
- Arandjelovic, O. (2016). Learnt quasi-transitive similarity for retrieval from large collections of faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4883–4892.
- Bartlett, M. S. (2001). Independent component representations for face recognition. In *Face Image Analysis by Unsupervised Learning*, pages 39–67. Springer.
- Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **19**(7), 711–720.
- Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, **35**(12), 2930–2940.
- Berg, T. L., Berg, A. C., Edwards, J., and Forsyth, D. A. (2004). Who’s in the picture? In *NIPS*.
- Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co.
- Bhattacharai, B., Sharma, G., and Jurie, F. (2016). Cp-mtml: Coupled projection multi-task metric learning for large scale face retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4226–4235.

- Brunelli, R. and Poggio, T. (1993). Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10), 1042–1052.
- Buades, A., Coll, B., and Morel, J.-M. (2005). A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, **4**(2), 490–530.
- Candes, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, **52**(12), 5406–5425.
- Candes, E. J., Romberg, J. K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, **59**(8), 1207–1223.
- Çeliktutan, O., Ulukaya, S., and Sankur, B. (2013). A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, **2013**(1), 13.
- Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, **14**(3), 462–467.
- Chowdhury, A. R. (2010). Fddb: Face detection data set and benchmark. <http://vis-www.cs.umass.edu/fddb/>.
- Conilione, P. and Wang, D. (2012). Fuzzy approach for semantic face image retrieval. *The Computer Journal*, **55**(9), 1130–1145.
- Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models—their training and application. *Computer vision and image understanding*, **61**(1), 38–59.
- Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 681–685.
- Dalal, N. (2005). INRIA person dataset. <http://pascal.inrialpes.fr/data/human/>.
- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.
- Dibeklioglu, H., Salah, A. A., and Akarun, L. (2008). 3d facial landmarking under expression, pose, and occlusion variations. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–6. IEEE.
- Ding, L. and Martinez, A. M. (2010). Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**(11), 2022–2038.

- Donoho, D. (2006). For most large underdetermined systems of linear equations the minimal l_1 -norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, **59**(6), 797–829.
- Duan, J., Pan, Z., Zhang, B., Liu, W., and Tai, X.-C. (2015). Fast algorithm for color texture image inpainting using the non-local ctv model. *Journal of Global Optimization*, **62**(4), 853–876.
- El Rube, I., Alajlan, N., Kamel, M., Ahmed, M., and Freeman, G. (2005). Efficient multiscale shape-based representation and retrieval. In *Image Analysis and Recognition*, pages 415–422. Springer.
- Felzenszwalb, P. F. and Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, **61**(1), 55–79.
- Fischler, M. A. and Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on computers*, (1), 67–92.
- Fisher, R. (1958). Statistical methods for research workers,(1925) oliver and boyd. *Edinburgh, England*.
- Freeman, W. T. and Roth, M. (1995). Orientation histograms for hand gesture recognition. In *International workshop on automatic face and gesture recognition*, volume 12, pages 296–301.
- Freeman, W. T., Tanaka, K.-i., Ohta, J., and Kyuma, K. (1996). Computer vision for computer games. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 100–105. IEEE.
- Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.
- Fritsch, F. N. and Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, **17**(2), 238–246.
- Frowd, C. D., Hancock, P. J., and Carson, D. (2004). Evofit: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on applied perception (TAP)*, **1**(1), 19–39.
- Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., and Zhao, D. (2008). The cas-peal large-scale chinese face database and baseline evaluations. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, **38**(1), 149–161.

- Gilboa, G. and Osher, S. (2008). Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, **7**(3), 1005–1028.
- Goldstein, T. and Osher, S. (2009). The split bregman method for l1-regularized problems. *SIAM Journal on Imaging Sciences*, **2**(2), 323–343.
- Gross, R. (2010). The CMU Multi-PIE Face Database. <http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html>.
- Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image Vision Computing*, **28**(5), 807–813.
- Gudivada, V. N., Raghavan, V. V., and Seetharaman, G. S. (1993). An approach to interactive retrieval in face image databases based on semantic attributes. In *Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas*.
- Heo, J., Kong, S. G., Abidi, B. R., Abidi, M., *et al.* (2004). Fusion of visual and thermal signatures with eyeglass removal for robust face recognition. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 122–122. IEEE.
- Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, **313**(5786), 504–507.
- Huang, J., Heisele, B., and Blanz, V. (2003). Component-based face recognition with 3d morphable models. In *International conference on audio-and video-based biometric person authentication*, pages 27–34. Springer.
- Huang, Y., Liu, Q., and Metaxas, D. (2007). A component based deformable model for generalized face alignment. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.
- Jain, V. and Learned-Miller, E. (2010). FDDB: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst.
- Jiang, X., Binkert, M., Achermann, B., and Bunke, H. (2000). Towards detection of glasses in facial images. *Pattern Analysis & Applications*, **3**(1), 9–18.
- Jing, Z., Mariani, R., and Wang, J. (2000). Glasses detection for face recognition using bayes rules. In *Advances in Multimodal InterfacesICMI 2000*, pages 127–134. Springer.
- Kahaner, D., Moler, C., and Nash, S. (1989). Numerical methods and software. *Englewood Cliffs: Prentice Hall, 1989*, **1**.

- Kan, M., Shan, S., Chang, H., and Chen, X. (2014). Stacked progressive auto-encoders (spae) for face recognition across poses. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1883–1890. IEEE.
- Karczmarek, P., Kiersztyn, A., Rutka, P., and Pedrycz, W. (2015). Linguistic descriptors in face recognition: A literature survey and the perspectives of future development. In *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2015*, pages 98–103. IEEE.
- Kasinski, A., Florek, A., and Schmidt, A. (2008). The PUT face database. *Image Processing and Communications*, **13**(3-4), 59–64.
- Kelly, G. A. (1955). *The psychology of personal constructs. Volume 1: A theory of personality*. WW Norton and Company.
- Kelly, G. A. (1969). A mathematical approach to psychology. *Clinical psychology and personality: The selected papers of George Kelly*, pages 94–113.
- Kendall, M. and Stuart, A. (1979). The advanced theory of statistics-volume 2 inference and relation-ship, (london: Charles griffin). *Kendall4The Advanced Theory of Statistics*, **2**.
- Koestinger, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2011a). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*.
- Koestinger, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2011b). Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. <https://lrs.icg.tugraz.at/research/aflw/>.
- Lanitis, A., Taylor, C. J., and Cootes, T. F. (1995). Automatic face identification system using flexible appearance models. *Image and vision computing*, **13**(5), 393–401.
- Lawrence, S., Giles, C. L., Tsoi, A. C., and Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, **8**(1), 98–113.
- Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. S. (2012). Interactive facial feature localization. In *Proceedings of the 12th European conference on Computer Vision - Volume Part III, ECCV’12*, pages 679–692, Berlin, Heidelberg. Springer-Verlag.
- Li, B., An, S., Liu, W., and Krishna, A. (2011). The mcf model: Utilizing multiple colors for face recognition. In *Image and Graphics (ICIG), 2011 Sixth International Conference on*, pages 1029–1034. IEEE.

- Li, B. Y., Mian, A. S., Liu, W., and Krishna, A. (2013). Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 186–192. IEEE.
- Li, S. Z. and Lu, J. (1999). Face recognition using the nearest feature line method. *IEEE transactions on neural networks*, **10**(2), 439–443.
- Li, Y., Wang, R., Huang, Z., Shan, S., and Chen, X. (2015). Face video retrieval with image query via hashing across euclidean space and riemannian manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4758–4767.
- Liu, C. and Wechsler, H. (2000). Evolutionary pursuit and its application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(6), 570–582.
- Liu, Y., Zhang, D., Lu, G., and Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, **40**(1), 262–282.
- Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, **60**(2), 91–110.
- MacTaggart, J. (2000). The proportions of the head and face. <http://www.artyfactory.com/portraits/pencil-portraits/proportions-of-a-head.html>.
- Mangold, C. (2007). A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies*, **2**(1), 23–34.
- Martínez, A. M. (1998). AR face database. <http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html>.
- Martínez, A. M. and Benavente, R. (1998). The AR face database. Technical Report 24, Computer Vision Center, Bellaterra.
- McConnell, R. K. (1986). Method of and apparatus for pattern recognition. US Patent 4,567,610.
- Mian, A. S. (2013). Databases. <http://staffhome.ecm.uwa.edu.au/~00053650/databases.html>.
- Milborrow, S. and Nicolls, F. (2008). Locating facial features with an extended active shape model. In *Computer Vision–ECCV 2008*, pages 504–513. Springer.
- Milborrow, S. and Nicolls, F. (2014). Active Shape Models with SIFT Descriptors and MARS. *VISAPP*. <http://www.milbo.users.sonic.net/stasm>.
- Moler, C. B. (2008). *Numerical Computing with MATLAB: Revised Reprint*. Siam.

- Nair, P. and Cavallaro, A. (2009). 3-d face detection, landmark localization, and registration using a point distribution model. *IEEE Transactions on multimedia*, **11**(4), 611–623.
- Nefian, A. V. and Hayes III, M. H. (1998). Hidden markov models for face recognition. *choice*, **1**, 6.
- Okada, K., Steffens, J., Maurer, T., Hong, H., Elagin, E., Neven, H., and von der Malsburg, C. (1998). The bochum/usc face recognition system and how it fared in the feret phase iii test. In *Face Recognition*, pages 186–205. Springer.
- OLIVEIRA JR, L. and Thomaz, C. (2006). Captura e alinhamento de imagens: Um banco de faces brasileiro. *Relatório de iniciação científica, Depto. Eng. Elétrica da FEI, São Bernardo do Campo, SP*, **10**.
- Papageorgiou, C. P., Oren, M., and Poggio, T. (1998). A general framework for object detection. In *Computer vision, 1998. sixth international conference on*, pages 555–562. IEEE.
- Pathirage, C. S. N., Li, L., Liu, W., and Zhang, M. (2015). Stacked face de-noising auto encoders for expression-robust face recognition. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, pages 1–8. IEEE.
- Penev, P. S. and Atick, J. J. (1996). Local feature analysis: A general statistical theory for object representation. *Network: computation in neural systems*, **7**(3), 477–500.
- Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). Numerical recipes in c (2d ed.; cambridge).
- Ren, Y., Li, Q., Liu, W., Li, L., and Yan, H. (N/A). Semantic characterization for eye shapes based on directional triangle-area curve clustering (under review). *EURASIP Journal on image and video processing*.
- Righi, G., Peissig, J. J., and Tarr, M. J. (2012). Recognizing disguised faces. *Visual Cognition*, **20**(2), 143–169.
- Schmidt, A. (2008). PUT face database. <https://biometrics.cie.put.poznan.pl/index.php?view=article&id=4>.
- Shan, S. (2008). CAS-PEAL face database. <http://www.jdl.ac.cn/peal/>.
- Sharon, Y., Wright, J., and Ma, Y. (2009). Minimum sum of distances estimator: robustness and stability. In *American Control Conference, 2009. ACC'09.*, pages 524–530. IEEE.

- Shen, J. and Chan, T. F. (2002). Mathematical models for local nontexture inpaintings. *SIAM Journal on Applied Mathematics*, **62**(3), 1019–1043.
- Sim, T., Baker, S., and Bsat, M. (2002). The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 46–51. IEEE.
- Sridharan, K. (2006). *Semantic face retrieval*. ProQuest.
- Štruc, V. and Pavešić, N. (2009). Gabor-based kernel partial-least-squares discrimination features for face recognition. *Informatica*, **20**(1), 115–138.
- Štruc, V. and Pavešić, N. (2011). Photometric normalization techniques for illumination invariance. *Advances in Face Image Analysis: Techniques and Technologies*, pages 279–300.
- Thomaz, C. E. (2006). FEI face database. <http://fei.edu.br/~cet/facedatabase.html>.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- Tkalcic, M. and Tasic, J. F. (2003). *Colour spaces: perceptual, historical and applicational background*, volume 1. IEEE.
- Torralba, A., Murphy, K. P., and Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **29**(5), 854–869.
- Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, **3**(1), 71–86.
- Valstar, M., Martinez, B., Binefa, X., and Pantic, M. (2010). Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE.
- Vaquero, D. A., Feris, R. S., Tran, D., Brown, L., Hampapur, A., and Turk, M. (2009). Attribute-based people search in surveillance environments. In *Applications of Computer Vision (WACV), 2009 Workshop on*, pages 1–8. IEEE.
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, **11**, 3371–3408.
- Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, **57**(2), 137–154.

- Wang, H., Li, S. Z., Wang, Y., and Zhang, J. (2004). Self quotient image for face recognition. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 2, pages 1397–1400. IEEE.
- Wang, Y., Duan, X., Liu, X., Wang, C., and Li, Z. (2016). Semantic description method for face features of larger chinese ethnic groups based on improved wm method. *Neurocomputing*, **175**, 515–528.
- Wang, Y.-K., Jang, J.-H., Tsai, L.-W., and Fan, K.-C. (2010). Improvement of face recognition by eyeglass removal. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on*, pages 228–231. IEEE.
- Wong, W. K. and Zhao, H. (2013). Eyeglasses removal of thermal image based on visible information. *Information Fusion*, **14**(2), 163–176.
- Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **31**(2), 210–227.
- Wu, C., Liu, C., Shum, H.-Y., Xy, Y.-Q., and Zhang, Z. (2004). Automatic eyeglasses removal from face images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **26**(3), 322–336.
- Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE.
- Yang, J. and Liu, C. (2008). A discriminant color space method for face representation and verification on a large-scale database. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE.
- Yang, J., Liu, C., and Zhang, L. (2010). Color space normalization: Enhancing the discriminating power of color spaces for face recognition. *Pattern Recognition*, **43**(4), 1454–1466.
- Yang, M., Kpalma, K., and Ronsin, J. (2008). A survey of shape feature extraction techniques. *Pattern recognition*, pages 43–90.
- Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE.
- Yin, L. (2008). Analyzing facial expressions in three dimensional space. http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html.

- Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE.
- Yin, L., Chen, X., Sun, Y., Worm, T., and Reale, M. (2008). A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference On*, pages 1–6. IEEE.
- Zhang, C. and Zhang, Z. (2010). A survey of recent advances in face detection.
- Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, **35**(4), 399–458.
- Zhu, X. and Ramanan, D. (2012a). Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR '12*, pages 2879–2886, Washington, DC, USA. IEEE Computer Society.
- Zhu, X. and Ramanan, D. (2012b). <http://www.ics.uci.edu/%7Exzhu/face/>.
- Zhu, X., Vondrick, C., Ramanan, D., and Fowlkes, C. (2012). Do we need more training data or better models for object detection?. In *BMVC*, volume 3, page 5. Citeseer.
- Zhu, Z., Luo, P., Wang, X., and Tang, X. (2013). Deep learning identity-preserving face space. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 113–120. IEEE.
- Zhu, Z., Luo, P., Wang, X., and Tang, X. (2014). Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.