

UNIVERSITY OF THE WESTERN CAPE

Robust Facial Expression Recognition in the Presence of Rotation and Partial Occlusion



A thesis submitted in fulfillment for the
degree of Master of Science

in the
Faculty of Science
Department of Computer Science

Supervisor: Mehrdad Ghaziasgar
Co-supervisor: James Connan

February 2014

Declaration



I, Diego Mushfieldt, declare that this thesis “Robust Facial Expression Recognition in the Presence of Rotation and Partial Occlusion” is my own work, that it has not been submitted before for any degree or assessment at any other university, and that all the sources I have used or quoted have been indicated and acknowledged by means of complete references.

Signature: 

Date: 26/02/2014

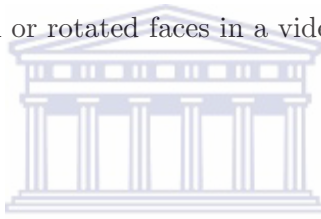
“A man provided with paper, pencil, and rubber, and subject to strict discipline, is in effect a universal machine.”

Alan Turing



Abstract

This research proposes an approach to recognizing facial expressions in the presence of rotations and partial occlusions of the face. The research is in the context of automatic machine translation of South African Sign Language (SASL) to English. The proposed method is able to accurately recognize frontal facial images at an average accuracy of 75%. It also achieves a high recognition accuracy of 70% for faces rotated to 60°. It was also shown that the method is able to continue to recognize facial expressions even in the presence of full occlusions of the eyes, mouth and left/right sides of the face. The accuracy was as high as 70% for occlusion of some areas. An additional finding was that both the left and the right sides of the face are required for recognition. As an addition, the foundation was laid for a fully automatic facial expression recognition system that can accurately segment frontal or rotated faces in a video sequence.



Keywords

Blender, Face Detection, Facial Expression Recognition, Haar Features, Local Binary Patterns, Morphological Operations, Occlusion, Rotation, Skin Detection, Support Vector Machine.

Acknowledgements

I would like to extend my gratitude to all my loved ones. Thank you for staying by my side throughout this journey. Thank you for your prayers, patience and constant motivation. In hard times, your words of encouragement uplifted me. To my supervisors, thank you for your positive feedback throughout the duration of this project. Your work ethic and leadership abilities are inspiring. Most importantly, thank you for your sarcasm/humour which always eased the tension and taught me not to take things too seriously. To Mr. Dodds, your passion for computer science is an inspiration to us all. To Mr. Brown, thank you for your technical advice. Finally, to my colleagues, Mr. Achmed, Mr. Nel and Mr. Frieslaar, thank you for your support throughout the years. It was indeed a pleasure and an honour to work with you.

Contents

Declaration of Authorship	i
Abstract	iii
Keywords	iii
Acknowledgements	iii
List of Figures	vii
List of Tables	x
Abbreviations	xii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Research Question	4
1.3 Summary of Research Objectives	4
1.4 Premises	5
1.5 Thesis Outline	6
2 Related Work	7
2.1 Motion-Based Methods	8
2.1.1 Feature Point Tracking	11
2.1.2 Dense Flow Tracking	15
2.2 Model-Based Methods	20
2.2.1 Active Appearance Models (AAMs)	20
2.2.2 Active Shape Models (ASMs)	26
2.3 Appearance-Based Methods	31
2.3.1 Local Binary Patterns (LBPs)	31
2.3.2 Gabor Wavelets	39
2.4 Summary	43
2.5 Conclusion	44



3	Image Processing in Appearance-Based Facial Expression Recognition	46
3.1	Face Segmentation Techniques	46
3.1.1	Face Detection	46
3.1.1.1	Haar-like Wavelets	47
3.1.1.2	Integral Image	48
3.1.1.3	AdaBoost Learning Algorithm	49
3.1.1.4	Constructing a Cascade of Weak Classifiers	49
3.1.1.5	Testing the Viola-Jones Face Detection Algorithm	50
3.1.2	Morphological Operations	50
3.1.2.1	Dilation	52
3.1.2.2	Erosion	52
3.1.3	Eye Detection	53
3.1.3.1	EyeMapC	53
3.1.3.2	EyeMapL	54
3.1.3.3	Eye Map	54
3.1.4	Skin Detection	55
3.1.4.1	RGB Colour Space	55
3.1.4.2	Normalized RGB Colour Space	56
3.1.4.3	HSV Colour Space	56
3.1.4.4	$Y C_b C_r$ Colour Space	57
3.1.4.5	TSL Colour Space	57
3.1.4.6	A Colour Space Conducive to Skin Detection?	58
3.1.4.7	Skin Model	59
3.2	Feature Extraction Using Local Binary Patterns	61
3.2.1	The Original LBP Operator	61
3.2.2	LBP Histograms	62
3.2.3	The Extended LBP Operator	63
3.2.4	Uniform and Non-Uniform Patterns	63
3.3	Support Vector Machines	64
3.3.1	The Optimal Hyperplane	65
3.3.2	Classifying Non-linear Problems	68
3.3.3	Kernel Functions	70
3.3.4	Multi-Class SVM Approaches	71
3.3.4.1	One-against-All	71
3.3.4.2	One-against-One	71
3.3.4.3	Directed Acyclic Graph	72
3.4	Summary	72
4	System Design and Implementation	74
4.1	Face Segmentation	75
4.1.1	Frontal Face Segmentation	76
4.1.2	Rotated Face Segmentation	78
4.2	Feature Extraction	78
4.3	Training and Testing Phases	82
4.3.1	Training Set	82
4.3.2	Training Phase	84
4.3.3	Testing Phase	85

4.4	Simulating Occlusion	85
4.4.1	Summary	86
5	Experimental Results and Analysis	88
5.1	Face Segmentation Experiment	89
5.1.1	Data Set	89
5.1.2	Experimental Procedure	90
5.1.3	Criterion for an Accurately Segmented Face	90
5.1.4	Results and Analysis	91
5.2	Feature Vector and SVM Optimization	93
5.2.1	Experimental Procedure	93
5.2.2	Results and Analysis	94
5.3	Facial Expression Recognition Accuracy Testing	95
5.3.1	Criterion for a Correctly Recognized Facial Expression	96
5.3.2	Frontal Facial Images	96
5.3.2.1	Experimental Procedure	96
5.3.2.2	Results and Analysis	96
5.3.3	Rotated Facial Images	102
5.3.3.1	Experimental Procedure	102
5.3.3.2	Results and Analysis	103
5.3.4	Frontal Occluded Facial Images	108
5.3.4.1	Experimental Procedure	108
5.3.4.2	Results and Analysis	109
5.3.5	Rotated Occluded Facial Images	114
5.3.5.1	Experimental Procedure	114
5.3.5.2	Results and Analysis	115
5.4	Summary and Conclusions	118
6	Conclusion	122
6.1	Directions for Future Work	123
6.1.1	Selecting a Suitable Database	123
6.1.2	Fully Automatic Systems	123
6.1.3	Comparing the Effects of Partial Occlusions and Rotations of the Face Using LBPs and Gabor Filters	124
6.2	Concluding Remarks	124
A	Additional Test Results	125
	Bibliography	134

List of Figures

2.1	Assumptions behind the Lucas-Kanade algorithm [12].	8
2.2	An example of the implemented Lucas-Kanade technique [12].	9
2.3	Pyramid Lucas-Kanade optical flow [12].	10
2.4	Physical markers manually placed on the face[113].	10
2.5	Mean Pearson correlation coefficients for a comparison between the MSRA method and the L-K optical flow algorithm [113].	11
2.6	The standard facial model [19].	12
2.7	Displacement of feature points [19].	13
2.8	The 12 facial feature points [11].	14
2.9	The six local feature vectors [11].	14
2.10	Occlusion results [11].	15
2.11	The bounding box and individual point displacements [95].	16
2.12	The six feature vectors representing the average displacement in each of the sub-regions [95].	16
2.13	The architecture of the system [37].	17
2.14	Velocity vectors in each frame [37].	17
2.15	Facial zones [37].	18
2.16	Synthesis module controls [37].	18
2.17	Flow fields indicating motion [89].	19
2.18	An example of a video sequence [89].	19
2.19	FER accuracy of the six HMMs across the video sequence [89].	20
2.20	An active appearance facial model [110].	21
2.21	A high-level breakdown of the system [25].	22
2.22	[25].	22
2.23	Independent base images [116].	25
2.24	AAM with manually labelled facial landmarks [116].	25
2.25	Wilhelm <i>et al.</i> 's FER results [116].	26
2.26	Manually annotated facial landmarks for ASMs [18].	26
2.27	Example of a shape model [18].	27
2.28	Lip tracking results using ASMs [71].	28
2.29	The shape model with 58 landmark points [18].	28
2.30	Error comparison of Chang <i>et al.</i> 's method and the ASM tracker [18].	29
2.31	The ASM facial model [96].	29
2.32	[96].	30
2.33	The original LBP operator.	32
2.34	Original (top row) and preprocessed (bottom row) images [34].	32
2.35	Frontal facial images [77].	34
2.36	Rotated facial images [77].	34

2.37	BU-3DFE results for the rotated view [77].	36
2.38	Six expressions in the multi-pie database [77].	36
2.39	An expression captured at different viewing angles [77].	37
2.40	Gabor kernels [27].	39
2.41	Real component of the Gabor representation of a facial image [27].	40
2.42	Magnitude of the Gabor representation of a facial image [27].	40
2.43	System overview [61].	40
2.44	Basis images [61].	41
2.45	Grids for each expression [61].	41
2.46	Partial occlusions of the face [61].	41
2.47	Gabor filters used by Liu and Wang [68].	42
3.1	Haar-like features.	47
3.2	48
3.3	Optimization of the Haar-like feature using the integral image.	49
3.4	AdaBoost feature selection.	49
3.5	Rejecting regions in an image.	50
3.6	Positive faces detected by the Viola-Jones algorithm [111].	51
3.7	Examples of various structuring elements [35].	51
3.8	The construction of <i>EyeMapC</i>	53
3.9	The construction of the final eye map.	54
3.10	Original colour image.	59
3.11	Skin image.	60
3.12	The original LBP operator.	61
3.13	Examples of texture primitives [97].	62
3.14	Concatenating region histograms into one single, spatially enhanced histogram [97].	62
3.15	Examples of texture primitives [97].	63
3.16	64
3.17	Decision boundary for the linear classification case.	65
3.18	The decision boundary in higher-dimensional space.	66
3.19	The decision hyperplane.	66
3.20	The optimal hyperplane separating two classes with a maximum margin [85].	67
3.21	Data points which are not linearly separable.	69
3.22	Directed Acyclic Graph of a 4-class problem. At each node a class is rejected until a single class remains.	72
4.1	FER framework.	74
4.2	High-level design of the algorithm.	75
4.3	Locating the nose.	76
4.4	Detecting the face.	76
4.5	An example of the normalization procedure.	77
4.6	Detecting the eye region.	77
4.7	Isolated frontal face.	78
4.8	Skin detection result for rotated images.	79
4.9	Skin image with morphological operations.	79

4.10	Detecting the contours of the face.	80
4.11	Isolated rotated face.	80
4.12	Applying the LBP operator.	81
4.13	Facial image divided into regions.	81
4.14	Example SVM training file.	82
4.15	The six prototypic facial expressions used from the BU-3DFE database [118].	83
4.16	SVM training procedure.	84
4.17	SVM prediction procedure.	85
4.18	System for frontal occluded images.	86
4.19	System for rotated occluded images.	86
5.1	An example of the five subjects each on a slightly different complex backgrounds.	90
5.2	Accurately segmented frontal and rotated faces for one subject.	91
5.3	The same histogram computed for the face and the detected object incorrectly perceived as the face.	92
5.4	An example of a case in which “Anger” was expressed similarly to “Disgust” [118].	98
5.5	An example of a case in which “Fear” was expressed similarly to “Happiness” [118].	98
5.6	FER accuracy per subject for frontal images.	99
5.7	Examples of cases in which “Sadness” was expressed similarly to “Anger” or a neutral expression.	100
5.8	Examples of cases in which “Fear” was expressed similarly to the neutral expression.	100
5.9	FER accuracy per expression for original frontal results and for the neutral cases removed.	102
5.10	FER accuracy per subject for rotated images.	105
5.11	FER accuracy per subject for rotated images.	106
5.12	Frontal and rotated images for Subject 24.	107
5.13	FER accuracy per expression for rotated images.	108
5.14	Simulated partial occlusion for frontal images for each region at (1/3), (2/3) and full occlusion.	108
5.15	Average accuracy across each progressive level of occlusion across all expressions.	112
5.16	FER accuracy per expression for frontal images fully occluded on the left and right sides.	113
5.17	FER accuracy per subject for unoccluded and fully occluded frontal images.	114
5.18	Simulated partially occluded rotated facial images.	115
5.19	Average accuracy across all expressions progressively occluding each region of rotated images.	117
5.20	FER accuracy per subject for frontal and rotated images fully occluded for each region.	118

List of Tables

2.1	Action units and corresponding facial expressions in the brow, eye and mouth regions [30].	12
2.2	Schweiger <i>et al.</i> 's results [95].	17
2.3	Number of samples in Datcu and Rothkrantz's dataset.	22
2.4	Confusion matrix using static images [25].	23
2.5	Confusion matrix using static images [25].	23
2.6	Confusion matrix for the FER experiment [108].	24
2.7	Action unit recognition results [108].	24
2.8	Results for the MBGC database [96].	30
2.9	Results for the Multi-PIE database [96].	30
2.10	BU-3DFE results for the frontal view [77].	35
2.11	Multi-Pie results for various angles [77].	37
2.12	Confusion matrix for facial expressions over all angles for LBP^{ms} features [77].	37
2.13	Confusion matrix for facial expressions over all angles for $LGBP$ features [77].	38
2.14	FER accuracy results of Shan <i>et al.</i> [97].	38
2.15	Occlusion results [61].	42
2.16	Gabor filter results [68].	43
5.1	Face segmentation accuracy for the tracking data set.	92
5.2	Optimized resolution and region size for frontal images.	94
5.3	Optimized resolution and region size for rotated images.	95
5.4	Confusion matrix for frontal FER accuracy.	97
5.5	Average Frontal FER accuracy.	97
5.6	System response for subjects with 3 out of 6 expression recognition and below.	99
5.7	The total number of misclassified cases for each expression and the number of images of each expression that resemble the neutral expression.	101
5.8	Results for rotated facial images.	103
5.9	Comparison of average FER accuracy using frontal and rotated faces.	103
5.10	System response for Subject 24.	106
5.11	Results for each region and level of occlusion for frontal images.	110
5.12	Results for each region and level of occlusion for rotated images.	115
5.13	Summary of the results obtained for the BU-3DFE database and the locally collected database.	121
A.1	System response for frontal images for the 40 subjects.	127

A.2 FER accuracy per subject for frontal images.	128
A.3 Assessment of the frontal data set to determine the expressions that resemble the neutral expression (“1”) and those that do not (“0”).	129
A.4 System response for rotated images for the 40 subjects.	130
A.5 FER accuracy per subject for rotated images.	131
A.6 Results for each region and level of occlusion for frontal images.	132
A.7 Results for each region and level of occlusion for rotated images.	133



Abbreviations

2D	2 Dimensional
3D	Three Dimensional
AAM	Active Appearance Model
ASM	Active Shape Model
AU	Action Unit
BU-3DFE	Binghamton University 3D Facial Expression Database
CCA	Connected Ccomponent Analysis
DAG	Directed Acyclic Graph
DNMF	Discriminant Non-negative Matrix Factorization
FACS	Facial Action Coding System
FSM	Finite State Machine
FPS	Frames Per Second
FER	Facial Expression Recognition
GHz	Gigahertz
GLVQ	Generalized Learning Vector Quantization
GMM	Gaussian Mixture Models
GPA	Generalized Procrustes Analysis
HMM	Hidden Markov Model
HSI	Hue Saturation Intensity
HSL	Hue Saturation Lightness
HSV	Hue Saturation Value
ICA	Independent Component Analysis
JAFFE	Japanese Female Facial Expression Database
kNN	k-Nearest-Neighbour
LBP	Local Binary Pattern
LibSVM	Library of Support Vector Machines

L-K	Lucas-Kanade
MBGC	Multiple Biometric Grand Challenge
MIDI	Musical Instrument Digital Interface
MLP	Multi Layer Perceptron
MSRA	Maximal Static Response Assay
NN	Neural Network
PCA	Principle Component Analysis
RBF	Radial Basis Function
RGB	Red Green Blue
SASL	South African Sign Language
SVM	Support Vector Machines
TSL	Tint Saturation Lightness



Chapter 1

Introduction

1.1 Background and Motivation

Verbal communication is an important tool that allows individuals to connect with each other by sharing and exchanging information and ideas. This important life skill is used on a daily basis in places such as schools, businesses and malls. It is the very fabric that unites societies, allowing them to function.

Deaf¹ people are severely marginalized in society as they are not able to fully participate in the exchange of verbal information. South Africa has a population of about 52.98 million people [54], of which a small minority of only 300 000 people are Deaf [40]. Communication and interaction between the Deaf and hearing is a daunting task. There are two main reasons for this: the Deaf community is the minority; and there are common misconceptions that the hearing have about the Deaf [52, 74].

Common misconceptions are: only a single sign language exists; sign languages are merely visual-gestural representations of spoken languages; linguistic studies can be applied to sign languages; and sign language sentences can be written using spoken words [52, 103]. Research [74] has shown that sign languages are fully fledged languages with entirely unique grammatical and syntactic structures, distinct from their spoken language counterparts. There are various different sign languages throughout the world with most countries having their own unique sign language [64]: British Sign Language (BSL), American Sign Language (ASL), Japanese Sign Language (JSL), South African Sign Language (SASL), among others.

South African Sign Language is the official language of the Deaf in South Africa and is recognized by the South African constitution as one of the 11 official languages [74].

¹The social group that are completely unable to communicate in spoken languages.

Although this is the case, the Deaf community still faces problems such as poor socio-economic opportunities and poor access to public and information services. Lotriet notes that there are “gross injustices” at South African courts and police stations which provide little to no access to expert interpretation services to the Deaf [69]. This impedes the development of Deaf communities. A temporary solution is to employ sign language interpreters. However, SASL interpreters are scarce and very costly [3, 13].

The SASL project [39] at the University of the Western Cape is in the process of developing a real-time machine translation system that can automatically translate between SASL and English. The translation between SASL and English involves two distinct processes: translation of SASL to English; and the translation of English to SASL. The procedures and technologies in each of these processes are varied. This research involves the first process – SASL to English translation. As part of this process, semantic information is extracted from a video consisting of a Deaf individual communicating in SASL using computer vision.

Research has shown that any sign language gesture can be characterized by five fundamental parameters [47, 65]: hand shape, hand orientation, hand motion, hand location and facial expressions. Research has been conducted by the SASL project towards the recognition of each of the five parameters. Li [65] developed a hand shape estimation system. Naidoo [81] and Rajah [93] developed gesture recognition systems based on hand motion recognition. Achmed [3] developed a hand location recognition system. Brown [13] developed an improved hand location system which focused on optimizing the accuracy and speed of Achmed’s system.

Facial expressions are a crucial component of sign language phrases as they provide conscious and subconscious feedback from the listener to the speaker through lexical, adverbial and syntactic information [114]. The mood and tonality of the phrase is expressed by means of facial expressions. Research has consistently shown that the focus of the eye-gaze of Deaf signers within a conversation is the facial region, specifically the region around the mouth [16, 78, 79].

Research has been conducted into Facial Expression Recognition (FER) by the SASL group. All such research has focused on the recognition of sub units of facial expressions as defined by the Facial Action Coding System (FACS). These sub units are called Action Units (AUs). The FACS defines key muscles in the face which can be moved to produce specific facial expressions. FER using FACS would require the tracking of these muscles and subsequently combining configurations of these muscles to describe larger-scale facial expressions. Whitehill [114] compared the effect of local versus global segmentation of the face using Haar features and the AdaBoost algorithm towards the recognition of AUs. Sheikh [99] analyzed the effect of AU recognition on noise degraded

images. Vadapalli [107] also developed an AU recognition system using Gabor filters for feature extraction and recurrent neural networks and Support Vector Machines (SVMs) for classification. All of these research projects proved highly successful in recognizing AUs.

This research diversifies research at the group in two ways: it focuses on the recognition of facial expressions as a whole, and identifies a suitable technique for this purpose; and it focuses on achieving accurate facial expression recognition in the presence of rotations and partial occlusions of the face, which has not been carried out in the group. Examples of whole expressions are “Happy”, “Sad”, “Disgust” etc.

A powerful feature extraction technique for facial expression recognition is the relatively new Local Binary Pattern (LBP) operator [97]. The next chapter details the most prominent techniques used to recognize facial expressions, with a focus on recognizing facial expressions as a whole, and justifies the selection of this operator in this research. Several variants of this operator have been used for facial feature extraction. These operators are introduced in the next chapter and explained in detail in Chapter 3.

Additionally, research has been conducted into the recognition of facial expressions with the head in rotated positions using this operator. The rotations referred to are yaw rotations of the face along the vertical axis of the signer’s spine. However, no research has been conducted into the recognition of facial expressions in the presence of occlusions of the face using this operator. It is, therefore, also unclear how a combination of a rotated and partially occluded face can affect the recognition accuracy using the operator. This research focuses on investigating this question.

Unlike some other sign languages like American Sign Language (ASL), there is an acute shortage of SASL information and data sets. Recently, the Fulton School for the Deaf released a SASL dictionary that contains 732 of the most common SASL phrases, which, to our knowledge, is the only formally available data set at the present time. There is still no SASL phrase image or video data set in existence. There are, however, a number of extensive facial expression databases in existence such as the Binghamton University 3D Facial Expression (BU-3DFE) data set [118]. The majority of these data sets provide videos of subjects performing six expressions called the “prototypic expressions”. These are: “Happy”, “Sad”, “Disgust”, “Fear”, “Surprise” and “Anger”. Analyzing the Fulton School for the Deaf dictionary revealed that all of these expressions are used to express emotions in a variety of SASL phrases. Therefore, there is overlap. Therefore, this research focuses on recognizing these prototypic expressions and makes use of the data set mentioned to train and test the system.

While the focus of this research is the recognition of facial expressions, it additionally attempts to lay the foundation towards another area of interest to the group: a fully automatic facial expression recognition system. A major component of fully automatic FER strategies is the ability to accurately isolate the face in query images [5]. Once the face has been isolated, relevant features can be extracted for the recognition of facial expressions. The majority of research in the field requires some form of manual intervention during the facial segmentation procedure. This includes placing landmarks on key facial locations, manually segmenting the face etc. The problem of accurate segmentation is further complicated in cases when the face is rotated, partially occluded or both. Two other factors that can contribute to complexity of the problem include a complex background and varied skin tones, as is the case in the South African context. Producing a fully automatic face segmentation system that is robust to both rotations and partial occlusions of the face is beyond the scope of this research. However, towards this goal, this research proposes a fully automatic face segmentation strategy in the presence of rotations of the face, and that is robust to complex backgrounds and varied user skin tones.

The Viola-Jones algorithm [111] is a robust method used to isolate faces in images. However, rotations are known to severely affect the accuracy of the algorithm [62, 66, 91]. Rotations of the face may be common while performing sign language gestures [47]. Therefore, a method that can accurately segment the face regardless of the angle and skin tone of the user, and on a complex background, is proposed in this research.

1.2 Research Question

The following research questions are specified based on the previous section:

1. “Can the proposed face segmentation strategy accurately segment the face in facial images with varied skin tone, in the presence of rotations and on a complex background?”
2. “Can whole facial expressions be recognized at a high accuracy using the LBP operator in the presence of rotations and partial occlusions of the face?”

1.3 Summary of Research Objectives

1. Lay the foundation of a fully automatic face segmentation strategy for the SASL group which does not require any manual intervention and can accurately segment

the face in frontal and rotated positions. The strategy should also be robust to variations in skin colour and a complex background.

2. Implement a FER strategy that is robust to rotations and partial occlusions of the face, variations in skin colour and a complex background.
3. Use the fully automatic face segmentation strategy to seamlessly investigate the accuracy of the proposed FER strategy for occluded frontal and rotated faces.
4. Simulate occlusions of the face and investigate the effects of various types and levels of facial occlusion on the recognition accuracy of the FER strategy for both frontal and rotated faces.

1.4 Premises

- It is assumed that the first frame of sign language video to be used in training and testing will consist of the signer facing the web camera. This assumption is justified since a conversation generally starts with two conversational partners facing each other.
- It is assumed that the signer will stand in front of an arbitrary background and in natural lighting conditions. This is justified since the SASL project requires the most natural setting.

1.5 Thesis Outline

The remainder of the thesis is arranged as follows:

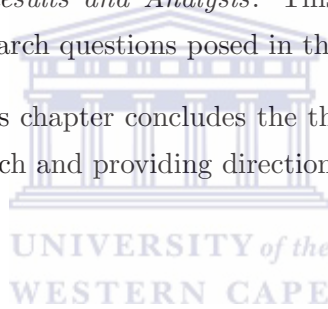
Chapter 2: *Related Work:* This chapter reviews existing literature in the field of facial expression recognition. An overview of each study is produced with much focus on the feature extraction techniques as it is a crucial step for facial expression recognition.

Chapter 3: *Image Processing in Appearance-Based Facial Expression Recognition:* This chapter provides details into the methods and algorithms used in the proposed system for the recognition of facial expressions.

Chapter 4: *Design and Implementation of the Robust Facial Expression Recognition in the Presence of Rotation and Occlusion System:* This chapter discusses the implementation of the proposed framework of the facial expression recognition system.

Chapter 5: *Experimental Results and Analysis:* This chapter discusses the testing carried out to answer the research questions posed in this chapter.

Chapter 6: *Conclusion:* This chapter concludes the thesis, highlighting the contributions made towards the research and providing directions for future work.



Chapter 2

Related Work

A generic framework for automatic facial expression recognition (FER) generally consists of three major components, namely: face detection, facial feature extraction and FER. Face detection is the process of locating and segmenting the face in each frame of the video sequence. Feature extraction is the process of analyzing the motion or texture properties within the facial region and extracting semantic information pertaining to the facial expression. FER is the process of determining the facial expression class corresponding to the extracted feature set.

This chapter presents a detailed survey on FER. Various researchers have mixed and matched different combinations of face detection, facial feature extraction and FER techniques. Therefore, it is not possible to categorize the studies according to all three components. This chapter categorizes the studies according to the facial feature extraction method used in each study as this has been argued to be the most important factor affecting the recognition accuracy [14, 77]. This is evidenced by the fact that the feature extraction methods in all such studies is explained, but the face segmentation and FER methods are not mentioned in a number of cases. Where possible, an explanation of the face segmentation and FER strategy of each study will be detailed.

Facial feature extraction methods can generally be sub-divided into three categories: motion-based methods; model-based methods; and appearance-based methods. Sections 2.1 , 2.2 and 2.3 provide details of each of these methods as well as the studies that have implemented them.

2.1 Motion-Based Methods

Motion-based methods associate a displacement measure with each pixel in the frame. The displacement measure provides information about the motion of various facial regions. This information can be used to characterize and determine the facial expression class associated with the facial motion. The large majority of motion-based FER systems use the Lucas-Kanade (L-K) optical flow algorithm [12]. The algorithm has been shown to be accurate in controlled conditions. However, the performance of this algorithm is sensitive to colour and intensity variations of the face. The more uniform and distinct the colour of the face, the higher the accuracy of the algorithm.

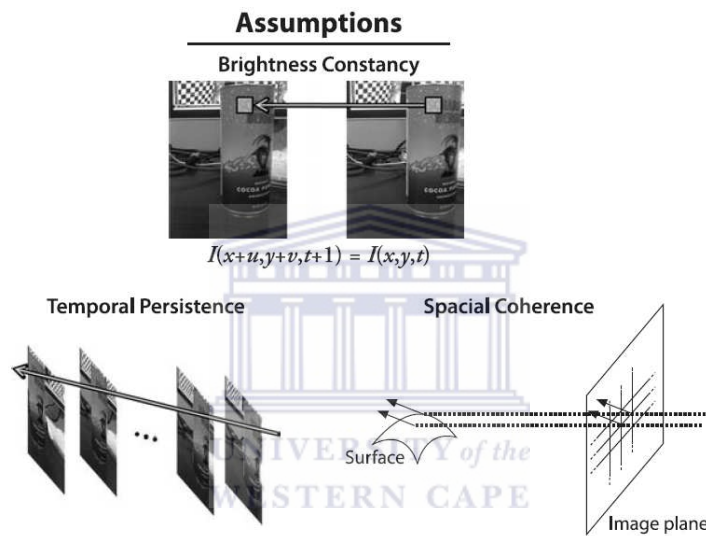


FIGURE 2.1: Assumptions behind the Lucas-Kanade algorithm [12].

The most popular feature point tracking technique is the Lucas-Kanade (L-K) sparse optical flow algorithm [12]. This algorithm rests on three key assumptions:

1. Brightness constancy – a pixel from the image of an object in the scene does not change in appearance as it moves from one frame to the next.
2. Temporal persistence – the image motion of a surface patch changes gradually in time.
3. Spatial coherence – neighbouring points in a scene belong to the same surface, have similar motion and project to nearby points on the image plane.

These assumptions are illustrated in Figure 2.1. This is mathematically expressed as:

$$I(x + u, y + v, t + 1) = I(x, y, t) \quad (2.1)$$

where x and y are the coordinates of the tracked pixel in the image, u and v are the changes in the x - and y -coordinates, t is the time, and I is the intensity of the tracked pixel. Figure 2.2 illustrates this tracking technique, which searches for the location of the required intensity value in order to track it in consecutive frames. The target pixel(s) are required to be manually specified in the initial frame in order to initialize tracking.



FIGURE 2.2: An example of the implemented Lucas-Kanade technique [12].

Since noise easily affects the performance of this method, the L-K optical flow technique assumes small changes in location and constant flow of a tracked pixel in a local neighbourhood. Only local information that is derived from a search window of known size surrounding each optical flow point is required, because the algorithm is applied in a sparse context. The use of smaller window sizes can easily cause the method to lose track of the tracked pixel if it falls outside the window when the motion is too fast. Conversely, the use of larger window sizes introduces sensitivity to noise, which is a similar disadvantage as using no window size at all.

In order to determine the optimal window size, an enhanced L-K optical flow technique known as the pyramidal L-K optical flow algorithm was developed [70]. In this technique, an initial 3×3 pixel window size is used to track the motion of a pixel by one pixel around its current location. A pyramid of increasingly smaller resolution copies of the image is created to manage the problem of large motions in the tracked object.

Referring to Figure 2.3 which is an example of a possible pyramid generated by the technique, the original image is at the base of the pyramid, with progressively smaller resolution copies of the image placed higher up in the pyramid structure. The L-K optical flow technique is applied to the highest level of the pyramid first to obtain an approximation of the location of the key feature point. This location is used to initialize a search window in and apply the L-K optical flow to the next level down in the pyramid. This procedure is repeated until the exact location of the pixel is located in the original image at the base of the pyramid.

Wachtman *et al.* [113] determined the accuracy of the optical flow technique by comparing the tracking accuracy of the technique with a physical marking method known as the Maximal Static Response Assay (MSRA). Videos of nine subjects, two men and seven

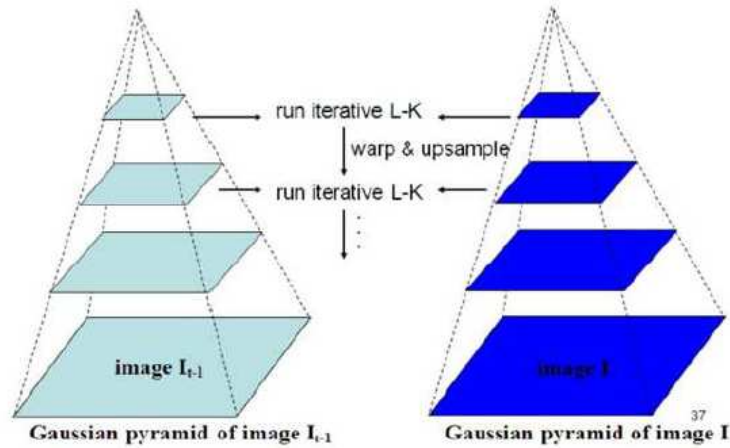


FIGURE 2.3: Pyramid Lucas-Kanade optical flow [12].

women from the Facial Nerve Centre at the University of Pittsburgh with an average age of 39 years, performing three expressions, “brow raise”, “eye closure” and “smile”, were recorded. In each case, subjects were asked to perform the expressions starting with the face in a relaxed state – the repose image – and ending with the peak of the expression – the peak image.

In the MSRA approach physical markers were manually placed on key locations on the face of each subject. The horizontal and vertical displacement of each point between the repose image and the peak image was manually computed. In the computer vision approach, L-K optical flow was used to automatically track the same points and compute the horizontal and vertical displacement of each point between the repose image and the peak image. Figure 2.4 illustrates the key points that were tracked using both techniques.

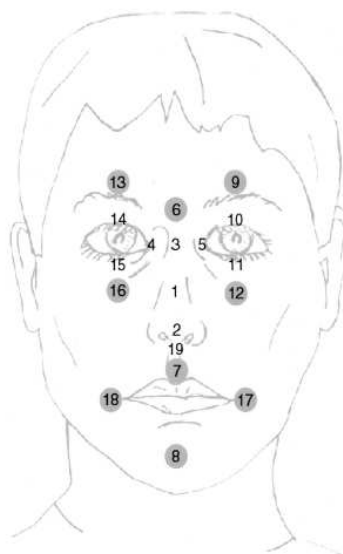


FIGURE 2.4: Physical markers manually placed on the face[113].

Pearson’s product-moment correlation was used to evaluate the consistency between the two techniques. Figure 2.5 summarizes the horizontal and vertical displacement correlation of the two techniques for each expression. The results indicate that the two systems are highly consistent with an average correlation ranging between 0.95 and 0.99. This indicates that the L-K optical flow algorithm is a very accurate tracking technique.

Action Task	Horizontal Displacement		Vertical Displacement	
	R*	SD	R*	SD
Brow raise	0.96	0.05	0.98	0.02
Eye closure	0.95	0.07	0.98	0.03
Smile	0.98	0.03	0.99	0.01

* Average correlation for 16 features for each action task. SD, standard deviation of the correlation coefficient; all correlation coefficients were significant, $p < 0.001$.

FIGURE 2.5: Mean Pearson correlation coefficients for a comparison between the MSRA method and the L-K optical flow algorithm [113].

Motion-based methods can be sub-divided into two types: methods that use feature point tracking and those that use dense flow tracking. Feature point tracking techniques use the displacement of a manual distribution of key facial landmarks in contrast to dense flow tracking techniques which use the displacement of a grid of points overlaid onto the facial region to characterize facial expressions. Sections 2.1.1 and 2.1.2 describe the studies that have applied feature point tracking and dense flow tracking, respectively, to FER.

2.1.1 Feature Point Tracking

Cohn *et al.* [19] used feature point tracking to recognize a set of specific action units (AUs) using the Facial Action Coding System (FACS) as a guideline. FACS is a system designed by Ekman [30] that primarily distinguishes very subtle facial features from each other. The system encodes the contraction or relaxation of specific muscles on the face into AUs. An action unit in this system is either a contraction or relaxation of a specific muscle in the face. Action units can be used individually or as a combination to describe facial expressions. The system describes 44 unique AUs that are able to represent all visible expressions. Examples of AU descriptions are illustrated in Table 2.1.

Three manually selected fiducial points are used to normalize the face in the image to overcome in-plane rotations, illustrated in Figure 2.6. No face segmentation strategy is required since the normalization points are manually specified. Additional feature points are manually selected on the face.

TABLE 2.1: Action units and corresponding facial expressions in the brow, eye and mouth regions [30].

Action Unit	Facial Expression
Brows	
AU 1+2	Inner and outer portions of the brows are raised.
AU 1+4	Medial portion of the eyebrow is raised and pulled together.
AU 4	Brows are lowered and drawn together.
Eyes	
AU 5	Upper eyelids are raised which produces a widening of the eyes.
AU 6	The lower eye and infra-orbital furrows are raised and deepened and the eye opening is narrowed.
AU 7	Lower eyelids are tightened, which narrows the eye opening.
Mouth	
AU 27	Mouth is stretched open and mandible extended.
AU 26	Lips are relaxed and parted; mandible lowered.
AU 25	Lips are relaxed and parted; mandible not lowered.
AU 12	Lip corners are pulled up and backward.
AU 12+25	AU 12 with mouth opening.
AU 20+25	Lips are parted, pulled back laterally, and may be slightly raised or pulled down.
AU 15+17	Lip corners are pulled down and stretched laterally (AU 15), and chin boss is raised, which pushes up the lower lip (AU 17).
AU 17+23+24	AU 17 and the lips are tightened, narrowed, and pressed together (AU 23+24).
AU 9+17±25	The infra-orbital triangle and centre of the upper lip are pulled upwards (AU 9) with AU 17. In 25% of cases, AU 9+17 occurred with AU 25.

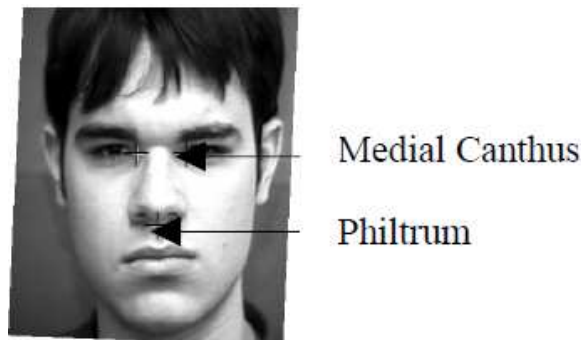


FIGURE 2.6: The standard facial model [19].

The Lucas-Kanade [70] algorithm is used to track the facial feature points. The displacement of each point is computed by subtracting its normalized position in the initial

frame from its normalized position in the current frame. Figure 2.7 illustrates the temporal displacement of these points for a facial expression. A 12-dimensional displacement

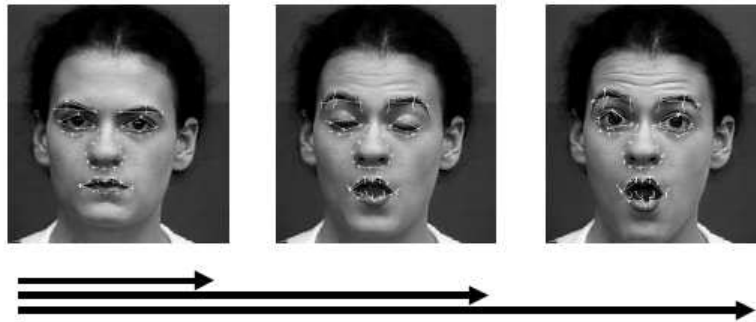


FIGURE 2.7: Displacement of feature points [19].

vector in the brow region, a 16-dimensional displacement vector in the eye region, a 12-dimensional displacement vector in the nose region and a 20-dimensional displacement vector in the mouth region is produced from the following horizontal and vertical flow vectors: six in the brow region; eight in the eye region; six in the nose region; and 10 in the mouth region. Feature point displacements between the initial and peak frames are used as predictors. For classification, separate group variance-covariance matrices are used.

A database consisting of 504 frontal image sequences containing 872 AUs from 100 subjects on a simple background were used. The dataset was randomly divided into training and cross-validation or test sets. The system achieved the following AU recognition accuracies: 92% for AUs in the brow region, 88% for AUs in the eye region and 83% for AUs in the nose and mouth regions.

Bourel *et al.* [11] investigated FER of four expressions, “Anger”, “Joy”, “Sadness” and “Surprise”, in the presence of partial occlusions of the face. The approach used a combination of L-K optical flow algorithm, a feature extractor, a group of k-nearest-neighbour (kNN) classifiers and a fusion module which combines the local classifiers.

A total of 12 facial feature points are manually specified around the following local regions of the face: three points on each eyebrow, one point on each nostril and four points around the mouth. No segmentation of the face is required since the points are manually specified on the face. The two points in the nostril region are used as a reference for automatic recovery. The system takes as input a video sequence starting at the neutral expression and ending at the peak of the expression. In each video sequence the displacement of each point is computed. Figure 2.8 illustrates the tracking procedure.

Six local feature vectors, as depicted in Figure 2.9, are created from the feature points on the face: a1, a2, c1, c2, c3, c4, d1 and d2. Four parabolic coefficients are extracted from

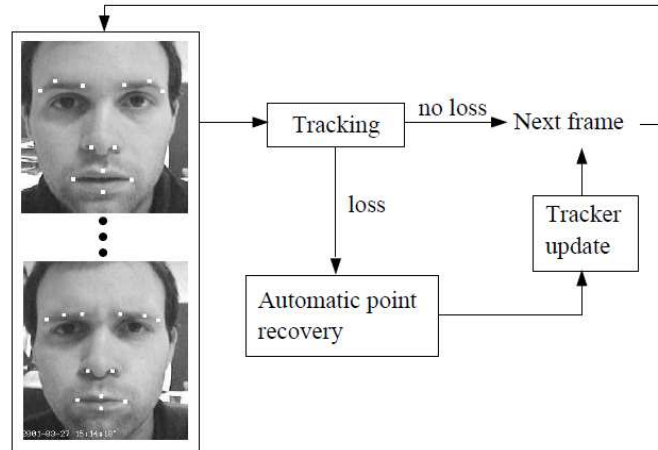


FIGURE 2.8: The 12 facial feature points [11].

the mouth region which model the shape of the mouth in each frame. Similarly, the brow region is also modelled using two coefficients which represent the angle of deformation of the eyebrows. The remaining two coefficients represent the distance between the eyebrows and nostrils. Six local feature vectors are constructed by taking the difference between the values of these coefficients in the current frame and those in the initial neutral frame.

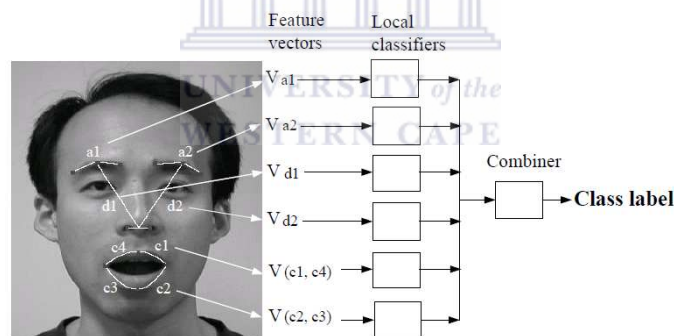


FIGURE 2.9: The six local feature vectors [11].

Each feature vector is fed into a local classifier. Each local classifier is a rank-weighted kNN classifier which produces a weighted score for each known expression class. This score is directly proportional to the rank of each nearest neighbour belonging to the class. The class of the first nearest neighbour will have the highest score followed by the class of the second and so on. The scores of all classes produced by all classifiers are summed. This yields a class-specific score for each expression class. The unknown pattern corresponds to the class with the highest score.

The accuracy of the FER strategy was tested under multiple occlusion settings: no occlusion, occlusion of the upper face, occlusion of the mouth and occlusion of sides of the face. Image sequences from the Cohn-Kanade database were used to train and test

the system. The Cohn-Kanade database [57] contains frontal images of subjects with varied skin tones on a simple background performing a variety of facial expressions. A total of 100 image sequences, 25 image sequences per facial expression class, from 30 subjects were used in experimentation. A leave-one-out cross-validation technique was used to test the system on the dataset.

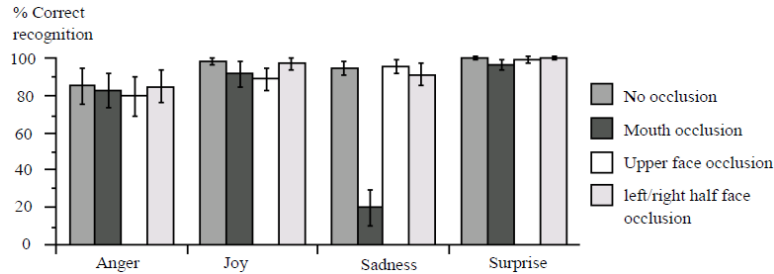


FIGURE 2.10: Occlusion results [11].

The results are displayed visually in Figure 2.10. The results indicate accuracies of 80% and above for all types of occlusion, with the exception of the “Sadness” expression with the mouth occluded, which registered the lowest accuracy. As per expectation, no occlusion of the face yielded the highest recognition accuracy. In the case of “Anger” and “Joy”, the recognition was mostly affected by occlusion of the upper face. This was not the case with “Sadness” and “Surprise” which were mostly affected by occlusion of the mouth. Occlusion of the left/right side of the face affected the recognition accuracy less than occlusion of the mouth and upper face in all expressions except for “Sadness”.

2.1.2 Dense Flow Tracking

Schweiger *et al.* [95] developed a framework for recognizing facial expressions. A manual face segmentation technique is used whereby the face in the first frame of a video sequence that is being analyzed is selected by drawing a bounding box around it. The box is carefully drawn such that it contains only the region from the top of the eyebrows to the bottom of the chin.

A grid of 64 equally spaced points is superimposed onto the facial image. The facial region is also divided into six sub-regions by means of one vertical line passing through the centre of the nose and two horizontal lines passing through the centres of the eyes and the mouth. Figure 2.11 illustrates the segmented face, the grid of points and the six sub-regions of the face.

The Lucas-Kanade algorithm [70] is used to track the flow of each point in the grid. A feature vector for each sub-region is computed which contains the average displacement for each sub-region in the grid. This results in a total of six feature vectors, the essence

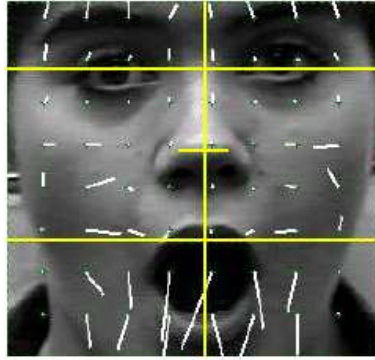


FIGURE 2.11: The bounding box and individual point displacements [95].

of which is illustrated in Figure 2.12.

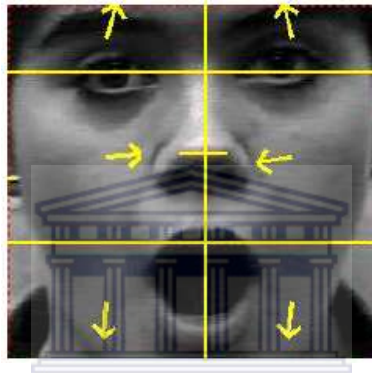


FIGURE 2.12: The six feature vectors representing the average displacement in each of the sub-regions [95].

The Fuzzy ARTMAP neural network architecture [15] is used to classify the facial expressions using the feature vectors as input. When the neural network receives a feature vector, the best-matching category is deduced by evaluating a distance measure against all category nodes. One neural network is trained for each of the six facial expressions.

The Cohn-Kanade facial expression database was used in the experimentation. The leave-one-out cross validation technique was used to obtain a recognition accuracy. The researchers state that the test set was inhomogeneous. Only a few test videos were available for “Fear” and “Disgust”. The results are summarized as a confusion matrix in Table 2.2. The last column of the table summarizes the total number of sequences used for each expression.

The results indicate a relatively high recognition accuracy for “Happiness”, “Sadness”, “Surprise” and “Anger”. The results for “Fear” and “Disgust” are insufficient to draw any conclusion about the recognition accuracies of these expressions.

Funk *et al.*'s system [37] associates facial movements with Musical Instrument Digital Interface (MIDI) notes that are sent to a sound synthesis module. The overview of the

TABLE 2.2: Schweiger *et al.*'s results [95].

	Happiness	Sadness	Surprise	Anger	Fear	Disgust	Total
Happiness	57	0	2	6	4	3	72
Sadness	3	26	4	8	2	0	43
Surprise	2	0	53	0	0	4	59
Anger	4	3	0	31	1	2	41
Fear	5	1	0	2	0	0	8
Disgust	5	0	0	2	0	3	10

system is depicted in Figure 2.13. For the purpose of this research, only the vision module

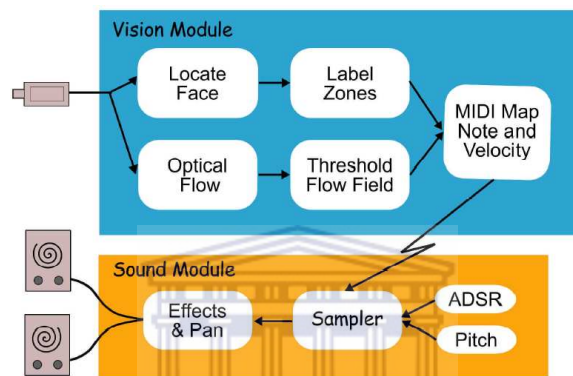


FIGURE 2.13: The architecture of the system [37].

will be taken into account. The system uses the Viola-Jones face detection algorithm [111] to automatically detect the face. A dense optical flow method known as “block matching” [43] is used to track facial movement. In this algorithm, a region of one frame of the video sequence is matched to a region of the same size in the subsequent frame of the video sequence. Matching is determined by calculating the sum of the absolute values of differences between pixels in the matching regions. The displacement of the block between the two frames results in velocity vectors, as illustrated in Figure 2.14.

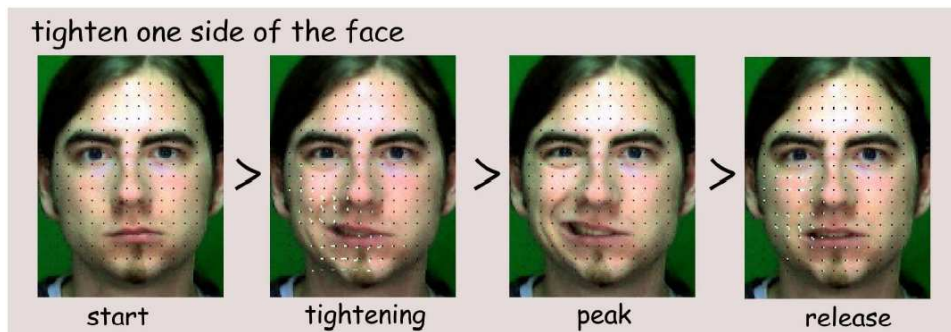


FIGURE 2.14: Velocity vectors in each frame [37].

Seven facial zones are approximated on the face which are used to label the motion vectors. The facial zones are illustrated in Figure 2.15. The regions are labelled as

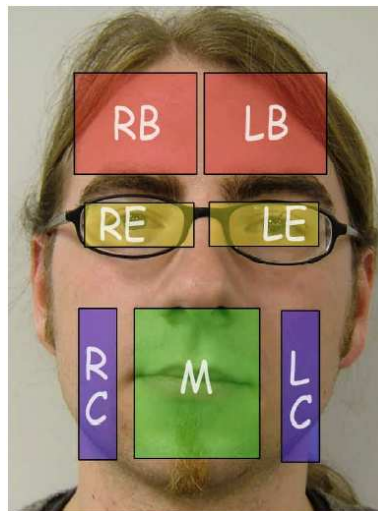


FIGURE 2.15: Facial zones [37].

follows: LB, RB – upper left and right eyebrow regions; LE, RE – left and right eyes; LC, RC – left and right cheeks; and M – the mouth. The vertical coordinate of each point on the grid determines the pitch of the generated MIDI notes and the magnitude of the flow vector determines the velocity. Figure 2.16 illustrates a screen shot of the synthesis control interface.

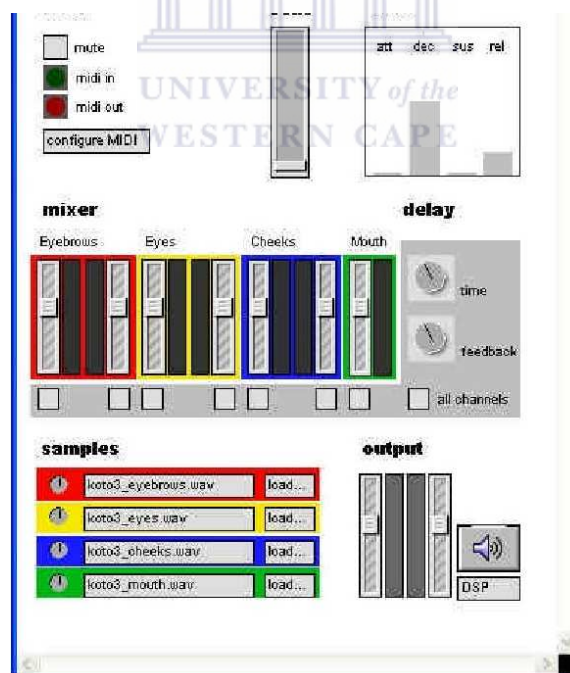


FIGURE 2.16: Synthesis module controls [37].

In order to control MIDI events, notes from each facial zone are sent on a separate MIDI channel, resulting in a total of 7 channels. The channels in the figure are colour-coded as red, yellow, blue and green, respectively. These are associated with the four facial zones: brows; eyes; cheeks; and mouth, in Figure 2.15. Each sample channel has the

stereo pan value set to its relative topography position corresponding to the topography of the face.

Otsuka and Ohya [89] developed a method of spotting segments in a video sequence that display facial expressions. Their approach uses a gradient-based optical flow algorithm [46] to estimate the motion around the right eye and mouth regions. The regions and their corresponding motion vector flow fields are depicted in Figure 2.17. The face detection procedure is not mentioned in the literature. A two-dimensional Fourier transform is

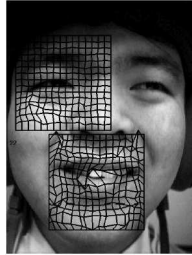


FIGURE 2.17: Flow fields indicating motion [89].

applied to the average velocity field and the lower-frequency coefficients are extracted as a 15-dimensional feature vector. The temporal sequence of the feature vector is mapped to its corresponding models which represent facial expressions.

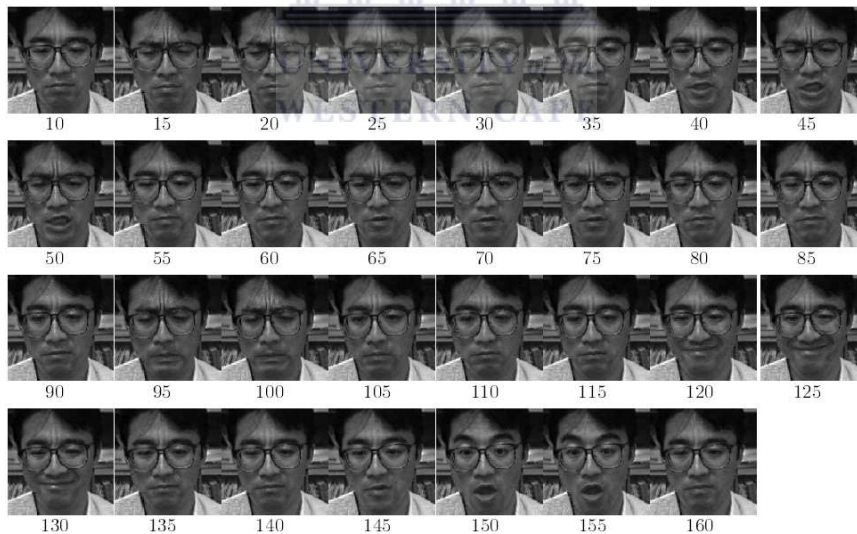


FIGURE 2.18: An example of a video sequence [89].

A Hidden Markov Model (HMM) [88] is used to recognize the six prototypic facial expressions. An HMM is a type of Finite State Machine (FSM) in which the state to be reached by a transition, as well as the vector produced by the transition, is non-deterministic. Each state corresponds to the conditions of facial muscles, namely: relaxed; contracting; and apex.

Two male subjects were instructed to perform the expressions starting with the neutral expression, progressing towards the peak of the expression in question, and ending at the neutral expression. Videos were recorded at a frame rate of 10 frames per second and each video consists of all the expressions displayed after each other, with two neutral frames inserted between different expressions, as depicted in Figure 2.18. Six prototypic facial expressions were displayed in an interval of 15 seconds. A neutral face was inserted between every two expressions. The HMM was trained and tested on the two male subjects.

The videos were recorded in a constrained environment. The experiment aimed at assessing the recognition rate of multiple expressions, as illustrated in Figure 2.19.

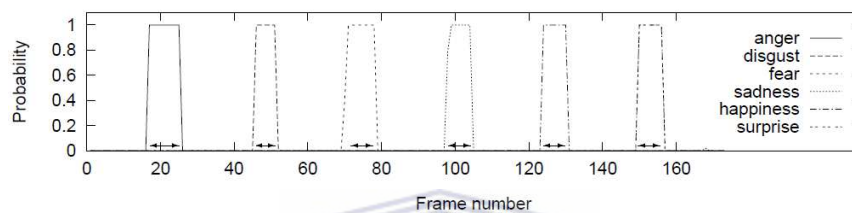


FIGURE 2.19: FER accuracy of the six HMMs across the video sequence [89].

The graph in the figure depicts the average recognition accuracy of the six HMMs each trained to recognize one of the prototypic expressions against time expressed as frame number in the video sequence. It is observed that in each interval in which one of the six expressions was performed, the recognition accuracy of the corresponding HMM peaked. It is observed that the system achieves a very high recognition accuracy.

2.2 Model-Based Methods

Model-based methods use statistical models to interpret facial images and provide a basis for the explanation of the appearance of the face using a set of model parameters. Model-based methods are divided into two categories, those that use active appearance models and those that use active shape models. Section 2.2.1 discusses active appearance models and Section 2.2.2 discusses active shape models. The details of these techniques are explained in the following subsections.

2.2.1 Active Appearance Models (AAMs)

An active appearance model is a powerful tool used for the extraction of a set of appearance parameters, from any unknown target face, coding a synthetic face similar to

the target with minimum error in texture [2]. Principle Component Analysis (PCA) is used to model both the shape and texture variations in the training set according to:

$$\mathbf{s}_i = \bar{\mathbf{s}} + Q_s \mathbf{c}_i \quad \text{and} \quad \mathbf{g}_i = \bar{\mathbf{g}} + Q_t \mathbf{c}_i \quad (2.2)$$

where Q_s and Q_t are truncated matrices describing the principle modes of combined appearance variations in the training set, and \mathbf{c}_i is a vector of appearance parameters simultaneously controlling the synthesized shape \mathbf{s}_i and texture \mathbf{g}_i . The mean shape $\bar{\mathbf{s}}$ and mean texture $\bar{\mathbf{g}}$ are computed on the aligned and normalized training faces.

For the purpose of model pose displacement, it is necessary to add to the appearance vector \mathbf{c}_i a pose vector \mathbf{p}_i which controls the scale, orientation and position of the synthesized face. Parameters \mathbf{c} and \mathbf{p} can automatically be adjusted by the active appearance model by minimizing a residual image $\mathbf{r}(\mathbf{c}, \mathbf{p})$ which is the texture difference between the synthesized face and the corresponding mask of the image that it covers. Figure 2.20 illustrates an example of applying the AAM to facial images.



FIGURE 2.20: An active appearance facial model [110].

The following studies have successfully implemented AAMs.

Datcu and Rothkrantz [25] compared the accuracy of FER between static images and video sequences using AAMs. The architecture of the system, from bottom-up, is depicted in Figure 2.21. The Viola-Jones face detection algorithm is used to detect segment the face in a frame. AAMs are used to model the face to obtain shape and texture data from it. The mean face shape illustrated in Figure 2.22(a) and mean texture depicted in Figure 2.22(b) modelled by the AAM account for the varied shapes and textures of the face from the training data. The final feature vector for static images consisted of 17 features pertaining to distances between key locations in the facial model. For video sequences, the variance occurring in each of the 17 features between the initial frame of the expression and the peak of the expression is computed and used as the 17-dimensional feature vector.

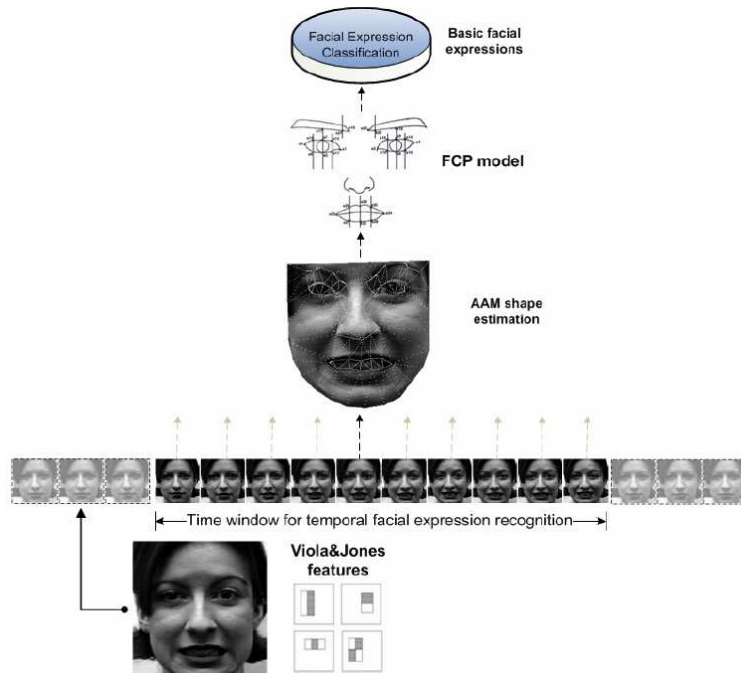


FIGURE 2.21: A high-level breakdown of the system [25].

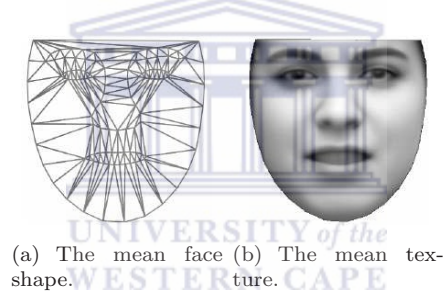


FIGURE 2.22: [25]

Finally, a Support Vector Machine (SVM) is used to classify the six prototypic facial expressions. The Cohn-Kanade database was used in the experimentation. The number of samples used for experimentation was different for each expression, and summarized in Table 2.3. Each of the samples in the table represent an entire sequence of images from a neutral expression to the peak of the expression and back to neutral. For static images, only the image representing the peak of the expression was selected, whereas for videos, the entire image sequence from neutral to the peak of the expression were used.

TABLE 2.3: Number of samples in Datcu and Rothkrantz's dataset.

Emotion	Sadness	Surprise	Anger	Fear	Disgust	Happy
Samples	92	105	30	84	56	107

Two-fold cross validation was used to train and test the system. The system achieved an average recognition accuracy of 80.02%, ranging from 72.64% to 84.70% for static

images, and an average recognition accuracy of 85.06%, ranging from 79.62% to 88.09% for video sequences. Table 2.4 and 2.5 illustrate the confusion matrices for static images and video sequences, respectively. The results indicate that the system registers a higher recognition accuracy with the use of video sequences as opposed to the use of static images.

TABLE 2.4: Confusion matrix using static images [25].

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	84.70	3.52	3.52	4.70	1.17	2.35
Surprise	12.38	83.80	0.95	0	0	2.85
Sadness	6.45	3.22	82.79	1.07	3.22	3.22
Anger	3.44	6.89	6.89	75.86	6.89	0
Disgust	0	0	7.14	10.71	80.35	1.78
Happy	7.54	8.49	2.83	3.77	4.71	72.64

Referring to Table 2.4, “Fear” registered the highest recognition accuracy of 84.7% and “Happy” registered the lowest recognition accuracy of 72.64%. It is interesting to note that “Surprise” was incorrectly registered as “Fear” in most incorrectly classified cases since these expressions are quite different.

TABLE 2.5: Confusion matrix using static images [25].

(%)	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	88.09	2.38	4.76	3.57	1.19	0
Surprise	0	88.67	2.83	8.49	0	0
Sadness	5.43	2.17	85.86	2.17	1.08	3.26
Anger	10.71	0	3.57	85.71	0	0
Disgust	5.35	5.35	3.57	1.78	82.14	1.78
Happy	4.62	0	7.40	2.77	5.55	79.62

Referring to Table 2.5, “Surprise” registered the highest recognition accuracy of 88.67% and, once again, “Happy” registered the lowest recognition accuracy of 79.62%. It should be noted that the recognition accuracy was higher in video sequences for every expression. It is also interesting to note that, in this case, “Surprise” was no longer misclassified as “Fear” at all.

Kuilenburg *et al.* [108] used a holistic implementation of an AAM to accurately recognize the six prototypic expressions and the neutral expression on static images. The constructed facial model is represented by an appearance vector which contains all the relevant information required to distinguish between expressions. No facial segmentation strategy is used as the system takes facial images on a simple background as input. A

three-layer feed-forward Neural Network (NN) is used for the classification of the expressions. A total of 116 neurons, 94 input neurons for the length of the appearance vector, 15 hidden neurons and 7 output neurons for each expression, were used in training the NN.

The Karolinska Directed Emotional Faces [72] database, which contains 980 facial images on a simple background, was used in the experimentation. The first experiment aimed at testing the accuracy at which the system can recognize facial expressions. The second experiment aimed at testing how accurately the system recognizes action units, using FACS as a guideline.

A total of 17 images were used for training and 963 images were used for testing. The training data consisted of 1512 appearance vectors. The testing data was not divided equally among the seven expressions. The results in Table 2.6 depict a confusion matrix of the first experiment and Table 2.7 summarizes the test results for recognizing action units.

TABLE 2.6: Confusion matrix for the FER experiment [108].

	Happy	Angry	Sad	Surprise	Scared	Disgust	Neutral
Happy	138	1	3	0	0	1	0
Angry	0	116	4	1	8	5	11
Sad	1	2	109	6	5	3	2
Surprise	0	1	19	128	2	0	1
Scared	0	3	2	0	115	3	1
Disgust	0	11	1	0	5	125	0
Neutral	1	0	1	0	3	0	125

For the first experiment the system achieved an average recognition accuracy of 89%. “Happy” registered the highest accuracy of 97% and “Sad” registered the lowest accuracy of 85%.

TABLE 2.7: Action unit recognition results [108].

Action Unit	01	02	04	05	06	07	09	12	15	17	20	23	24	25	27	Average
Accuracy (%)	86	88	81	86	81	89	93	83	89	86	84	83	83	90	89	86

Table 2.7 illustrates the results for the second experiment. For action unit recognition the system registered an average accuracy of 86%.

Wilhelm *et al.* [116] compared two models that classify facial expressions, age, gender and identity. The first model, Independent Component Analysis (ICA), is a description of facial images by their projection on independent base images. For this model, the

centres of the eyes are manually located and used as facial landmarks in facial images that contain a simple background. As such, the system requires no automatic face segmentation strategy. An observation matrix is computed using vectorized images as rows. Figure 2.23 illustrates the independent base images obtained as a result of applying ICA to the matrix.

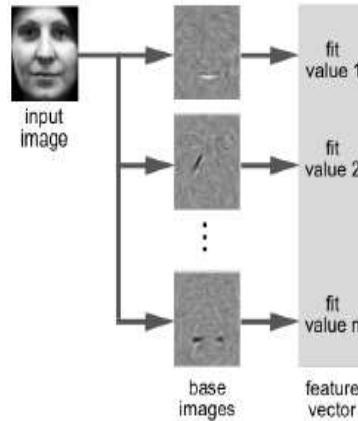


FIGURE 2.23: Independent base images [116].

The second model uses AAMs to model the shape and grey value variations of facial images. For this model, 116 facial landmark points were used along dominant outlines of the face. Facial landmarks are manually labelled on the face for normalization as illustrated in Figure 2.24.

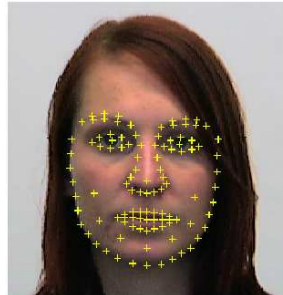


FIGURE 2.24: AAM with manually labelled facial landmarks [116].

The feature vectors were classified using various machine learning techniques such as Nearest Neighbours (NNs), Multi Layer Perceptron (MLP), Radial Basis Function (RBF) and Generalized Learning Vector Quantization (GLVQ). The database used to train and test the systems consisted of 30 subjects on a simple background performing the six prototypic expressions, as well as the neutral expression. Only static images, one per expression per subject, are contained in the database. Since the classification of age, gender and identity are beyond the scope of this research, only the test results for FER are presented. The results are illustrated in Figure 2.25.

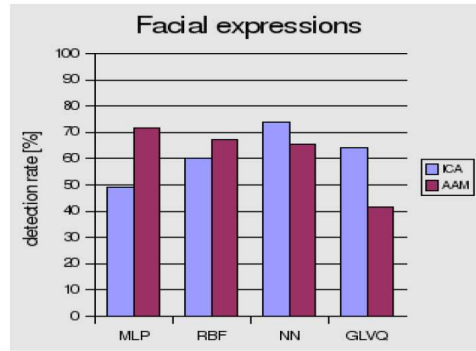


FIGURE 2.25: Wilhelm *et al.*'s FER results [116].

For the ICA method, the recognition accuracies were varied, with the NN achieving the best results and the MLP achieving the lowest accuracy. For the AAM method, the recognition accuracies are fairly consistent, with the MLP achieving the highest accuracy, except for the GLWQ classification technique which registered a much lower accuracy than the other three techniques.

2.2.2 Active Shape Models (ASMs)

Active Shape Models are statistical models of the shape of objects which iteratively deform to fit on an example of the object in a new image [20]. A statistical facial model is created from a training set of images consisting of manually annotated facial landmarks. An example of a landmarking scheme is illustrated in Figure 2.26.

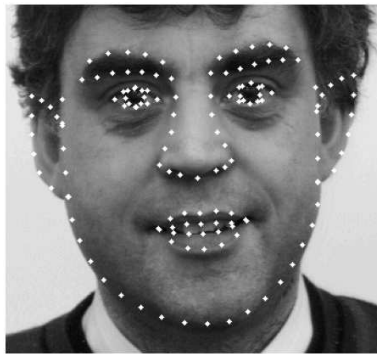


FIGURE 2.26: Manually annotated facial landmarks for ASMs [18].

For each facial image, the coordinates of all landmarks are stored as a vector, called a shape \mathbf{x} , in the form $\mathbf{x} = ((x_1, y_1), \dots, (x_N, y_N))^T$, where x_i and y_i are the coordinates of the i -th landmark and N is the number of landmarks used [96]. All shapes in the training set are aligned with each other using Generalized Procrustes Analysis (GPA) and the mean shape $\bar{\mathbf{x}}$ is the mean vector shape of these aligned shapes. Statistical

models of the grey level intensities of the region around each landmark are generated to build a subspace that spans the variations of the exemplar training images.

One-dimensional profiles are created by sampling the grey level intensities that lie around the lines of the face. These intensities are stored as a vector and normalized by replacing each element of the vector by its gradient and dividing the mean of the absolute values of its elements. The mean profile vector $\bar{\mathbf{g}}$ and the covariance matrix of all such vectors is denoted by \mathbf{S}_g .

For each landmark point, the mean profile vector and covariance matrix are generated. For generating two-dimensional profiles, the resulting matrix is vectorized, row-wise, and normalized by applying a sigmoid transform, with a shape constant q , to each element of the profile, g_i , to transform them into g'_i , as shown in Equation (2.3)

$$g'_i = \frac{g_i}{|g_i| + q} \quad (2.3)$$

An example of a facial model is illustrated in Figure 2.27.



FIGURE 2.27: Example of a shape model [18].

The following studies have successfully implemented ASMs.

Luettin *et al.* [71] applied ASMs to visual speech recognition. The inner and outer contours of the lips were used to create the model. In each frame the parameters that describe the shape lips are extracted and used as visual speech feature vectors. The temporal changes of these vectors were modelled by HMMs. The experiments aimed at testing word accuracy. The Tulips1 database contains grey level image sequences of the first four digits and each digit was spoken twice by 12 subjects on a simple background. Figure 2.28 depicts the visual results of locating and tracking the lips using ASMs.

The results indicate an accurate method for locating and tracking the lips using ASMs, even in cases when they extend beyond the boundaries of the image (2nd column and 2nd row of Figure 2.28 and 3rd column and 2nd row of the same figure). In terms of

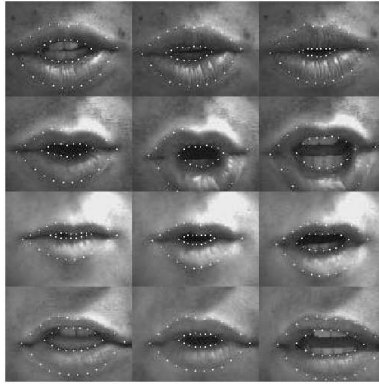


FIGURE 2.28: Lip tracking results using ASMs [71].

recognition accuracy, the system achieved an average recognition accuracy of 88.42% across all subjects and words.

Chang *et al.* [18] proposed a novel approach to recognizing facial expressions on a low-dimensional expression manifold. Facial deformations in a low-dimensional space are embedded using non-linear dimensionality reduction. Images lie in a very high-dimensional space. However, a class of images generated by latent variables lies on a manifold in this space. In human facial images, the latent variables may be the illumination, identity, pose and facial deformations. The facial model is manually selected in order to track facial deformation. A Gaussian Mixture Model (GMM) is applied to cluster data in the low-dimensional expression space in an off-line training phase. A specific ASM, defined by manually locating 58 facial landmarks, is trained for each cluster. The shape model is illustrated in Figure 2.29. The ICondensation algorithm, which is a probabilistic prediction model, is used for facial deformation tracking and recognition.

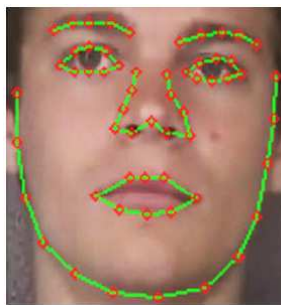


FIGURE 2.29: The shape model with 58 landmark points [18].

Two subjects were instructed to perform the six basic expressions in sequence seven times. Half of the data was used for training and the other for testing. The results in Figure 2.30 clearly indicate that Chang *et al.*'s method obtained a considerable improvement – less error – when compared to the traditional ASM method.

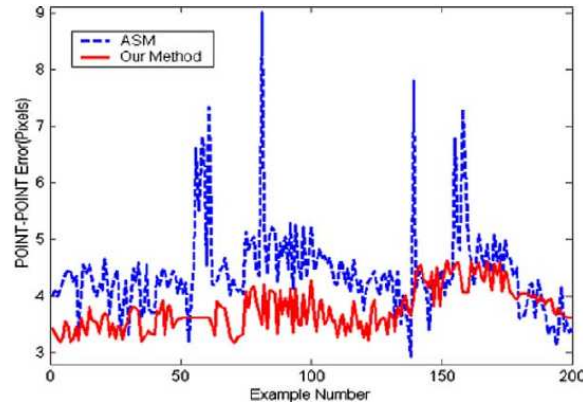


FIGURE 2.30: Error comparison of Chang *et al.*'s method and the ASM tracker [18].

Seshadri and Savvides [96] proposed an enhanced facial landmark optimization technique to improve the accuracy of ASMs. An optimal number of 79 landmark points were found to be sufficient to accurately model the face in order to carry out reasonable facial analysis, especially when dealing with facial expressions. The facial model is illustrated in Figure 2.31.

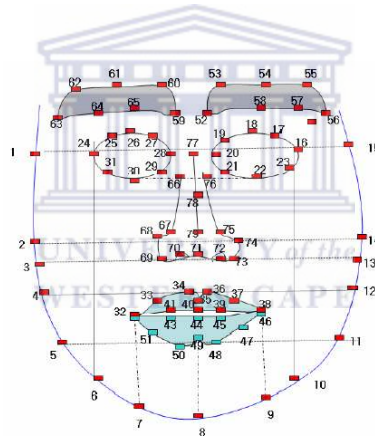


FIGURE 2.31: The ASM facial model [96].

A new metric which is the Mahalanobis distance between the original candidate profile patch and the reconstructed candidate profile patch was proposed. Therefore, the candidate patch with the lowest reconstruction error is deemed as the best fit. Figure 2.32(a) depicts one-dimensional profiles, used in traditional ASMs, which are constructed by sampling the grey level intensities along the lines known as whiskers [76]. Figure 2.32(b) illustrates how Seshadri and Savvides construct two-dimensional profiles by sampling a 13×13 square region around each landmark.

The experiment aimed at testing the modelling accuracy of the modified ASM method against the conventional ASM implementations on two datasets. The training set consisted of 500 images of 115 subjects from the NIST Multiple Biometric Grand Challenge (MBGC) database which contains a total of 10687 images of 570 subjects performing

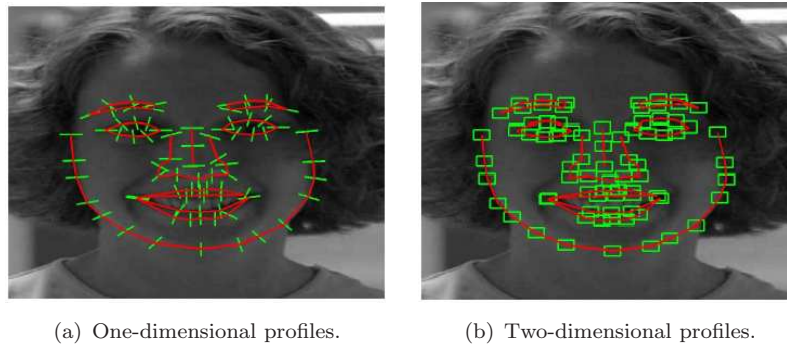


FIGURE 2.32: [96]

the six prototypic expressions. The testing sets consisted of 1500 images from the NIST Multiple Biometric Grand Challenge (MBGC) dataset and 500 images from the CMU Multi-Pie dataset.

Table 2.8 illustrates the test results for the MBGC dataset and Table 2.9 illustrates test results for the CMU Multi-Pie dataset. In both cases, the results indicate the error with which the modified ASM method models the face, as compared to classical ASMs.

TABLE 2.8: Results for the MBGC database [96].

Method Used	Average Fitting Error	Average Normalized Fitting Error
Classical ASM	12.101	3.501
Seshadri and Savvides' Implementation	6.582	1.908

Referring to Table 2.8 and Table 2.9, the results indicate that the proposed approach outperforms classical ASMs. The researchers state that, for the MBGC database, the proposed approach outperforms the classical ASM algorithm by slightly more than 45.5% on both the fitting error and average normalized fitting error. Similar results were obtained on the Multi-PIE database, for which the proposed approach is stated to be 30% more accurate than the conventional ASM on both the fitting error and average normalized fitting error.

TABLE 2.9: Results for the Multi-PIE database [96].

Method Used	Average Fitting Error	Average Normalized Fitting Error
Classical ASM	7.655	3.571
Seshadri and Savvides' Implementation	5.332	2.488

2.3 Appearance-Based Methods

Appearance-based methods rely on the dynamic shape and texture changes in the face in order to extract facial features. Appearance-based methods are divided into two types: Local Binary Patterns (LBPs) and Gabor wavelets. Section 2.3.1 discusses Local Binary Patterns and Section 2.3.2 discusses Gabor wavelets. The details of these techniques are explained in the following subsections.

2.3.1 Local Binary Patterns (LBPs)

Ojala *et al.* developed the original LBP operator [86]. The LBP operator is a good texture descriptor. Applied to an image, the operation results in a texture image that can be used in facial expression analysis. The operator is applied to a grey scale image and results in a grey scale texture image.

The operator is applied to each 3×3 pixel neighbourhood in the input image. For each pixel location, the following procedure is carried out:

1. The neighbouring pixels $\{f_P | P = 0, \dots, 7\}$ are assigned binary values by means of a threshold function which is relative to the value of the centre pixel f_c . The threshold function T is given by:

$$T(f_P, f_c) = \begin{cases} 1 & \text{if } f_P \geq f_c \\ 0 & \text{otherwise} \end{cases} \quad (2.4)$$

2. The values of the neighbouring pixels are taken to be an 8-bit binary number, with the binary value of the top left pixel as the left-most bit, and moving clockwise, until the neighbouring pixel on the left is encountered and taken to be the right-most bit of the binary number. This binary code is known as a local binary pattern. Note that the binary pattern can be generated by moving in the opposite direction as well and can start at any neighbour as long as the procedure remains consistent.
3. The binary pattern is converted to its decimal equivalent and assigned as the value of the pixel in a new LBP image in the location corresponding to the centre pixel in the original image.

This procedure is illustrated in Figure 2.33.

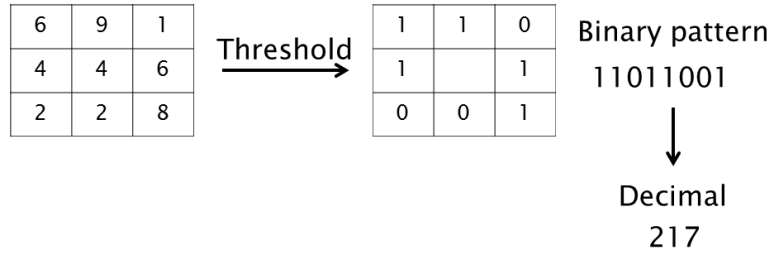


FIGURE 2.33: The original LBP operator.

The following studies have successfully implemented LBPs.

Feng *et al.* [34] used LBPs to recognize the six prototypic expressions as well as the neutral expression. Each image was preprocessed using the CSU Face Identification Evaluation System [10] resulting in segmented normalized facial images of size 150×128 pixels. Figure 2.34 depicts examples of images used from the JAFFE database before and after preprocessing.

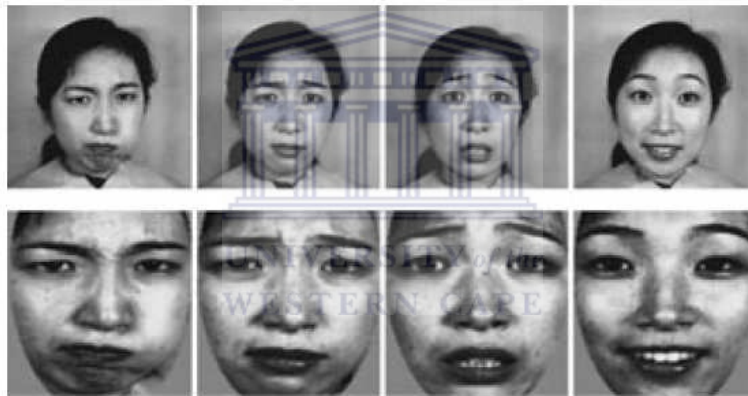


FIGURE 2.34: Original (top row) and preprocessed (bottom row) images [34].

The original LBP operator was applied to the image. The resulting image was divided into local regions of size 10×8 pixels. Local histograms of each region were computed and concatenated into one feature vector.

Classification was carried out by means of a linear programming technique. This technique generates a plane which minimizes an average sum of misclassified points belonging to two disjoint point sets. The seven-expression classification problem is divided into 21 two-class problems: Anger-Disgust, Anger-Fear, Disgust-Fear etc. For the training phase, 21 classifiers each corresponding to the 21 expression pairs are trained. For testing, the feature vector of each testing sample is fed into all the classifiers for recognition and a binary tree tournament scheme is used to resolve the multi-class classification problem.

The data used for experimentation was obtained from the JAFFE database [73]. A total of 213 images of 10 subjects on a simple background performing each of the expressions “three or four times” [34] were used. Ten-fold cross validation was used to train and test the system. An average recognition accuracy of 93.8% was obtained. It was noted that the “Fear” expression was problematic and difficult to recognize. When the “Fear” expression was disregarded, an average recognition accuracy of 94.6% was obtained. No further results are provided.

Moore and Bowden [77] investigated multi-view FER using LBPs. Several variants of the original LBP operator were used on facial images from the Binghamton University 3D Facial Expression (BU-3DFE) database [118] and the Multi-Pie database [42]. The aim of the study was to test the influence of pose on FER. Both databases contain facial images on a simple background. Various optimized LBP operators were used and compared. The following variants were considered: rotation invariant LBPs – LBP^{ri} ; standard uniform LBPs with a neighbourhood of eight pixels and a radius of one pixel – LBP^{u2} ; uniform rotation invariant LBPs – LBP^{riu2} ; uniform LBPs obtained from gradient magnitude images – LBP^{gm} ; multi-scale LBPs where the radius varies from one to eight pixels – LBP^{ms} ; and LBPs extracted from Gabor images using 40 different gabor kernels at different scales and orientations to compose gabor images – $LGBP$.

LBP^{ri} offers rotation invariance but has poor descriptive abilities. LBP^{u2} offers illumination invariance and is computationally efficient. LBP^{riu2} also offers rotation invariance and also has poor descriptive abilities. LBP^{gm} features encode the magnitude of local variation. LBP^{ms} allows for multi-scale analysis by encoding the micro features of the face as well as large-scale features at the structural level. $LGBP$ has a high feature vector dimensionality, but Gabor filters offer strong illumination invariance as well as powerful descriptive features.

In the approach, each of these operators is used to generate an LBP image. The resulting image is divided into 64 sub-regions using a grid of 8 columns and 8 rows. Histograms of each region are computed and concatenated to form a “spatially enhanced histogram” [77] which is used as a feature vector. A multi-class SVM is used to classify the resulting feature vector into classes representing each of the seven expressions.

The BU-3DFE database contains facial images at 4 intensities, ranging from neutral to the peak of the expression, for all 6 prototypic expressions as performed by 100 subjects. The subjects are of varied skin tone. The facial images are provided as 3D models that can be rotated to any angle. Figure 2.35 illustrates the frontal images from the BU-3DFE database and Figure 2.36 depicts images of a different subject at 5 different angles: 0° , 30° , 45° , 60° and 90° .

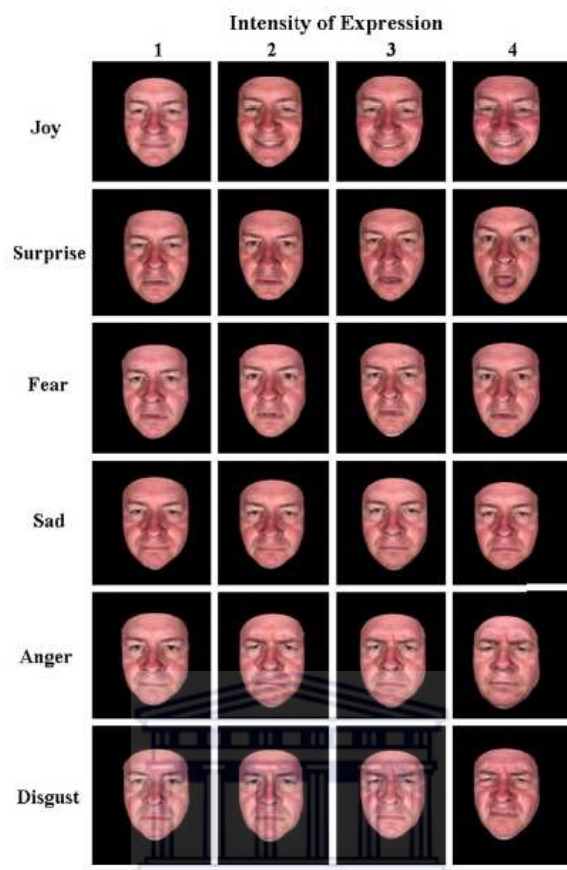


FIGURE 2.35: Frontal facial images [77].

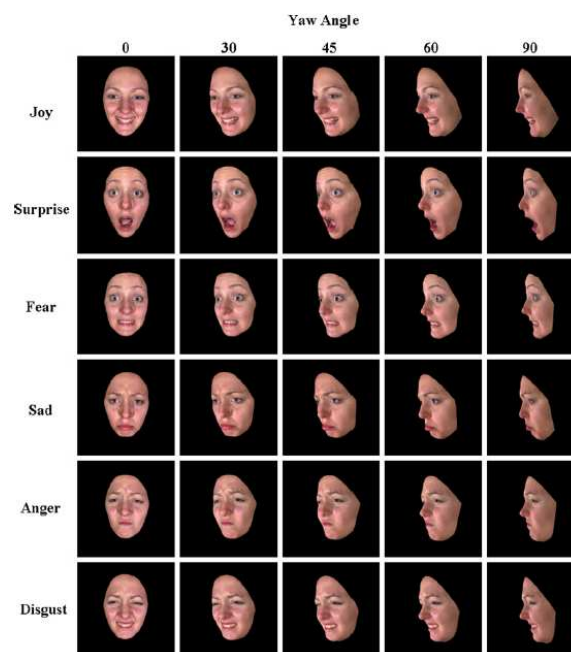


FIGURE 2.36: Rotated facial images [77].

No pre-processing was carried out on the facial images. An experiment was carried out to determine the FER accuracy of each LBP operator at the selected angles. Additionally, the effect of the resolution size of the original facial image on recognition accuracy was also investigated. Four randomly selected resolution sizes were used. The ten-fold cross-validation accuracy on 8000 images per class was computed and used as a comparative measure between each of the cases. Table 2.10 illustrates the average recognition accuracy over all subjects, expressions and angles at each resolution size.

TABLE 2.10: BU-3DFE results for the frontal view [77].

	32×44	44×62	64×88	80×110
LBP^{riu2}	47.28	46.12	46.31	46.32
LBP^{ri}	47.53	46.28	45.93	46.56
LBP^{gm}	52.91	51.49	53.2	53.29
LBP^{u2}	58.44	57.33	57.12	56.24
LBP^{ms}	62.41	62.9	64.98	65.02
$LGBP$	66.76	67.84	67.96	66.79

The results indicate that varied resolution sizes are suitable for different LBP operator variants and it thus necessary to determine the optimum resolution on a per-application basis. The three most accurate LBP operators were the $LGBP$, LBP^{u2} and LBP^{ms} .

Figure 2.37 summarizes the graphical results obtained per expression and viewing angle for these three operators. It can be observed that the recognition accuracy between the different expressions varied quite significantly for all three methods. However, the accuracy between different viewing angles appeared to be relatively consistent. Only the average accuracy of the two operators $LGBP$ and LBP^{ms} are provided, and these were 67.96% and 65.02%, respectively.

The multi-pie database contains images of 337 subjects of varied skin tone on a simple background performing five expressions, three of which are prototypic expressions – “smile”, “disgust” and “surprise” – with two other expressions “squint” and “scream”. The neutral expression is also included. Thirteen cameras are used to record each subject performing each expression at varied rotations in 15° intervals ranging from 0° – 180° . The orientation of the head of each subject is carefully controlled by a pre-configured head brace. Figure 2.38 illustrates the six expressions in the database and Figure 2.39 illustrates one expression captured at different viewing angles.

An experiment was carried out to determine the recognition accuracy of the two best performing operators from the previous experiment: $LGBP$ and LBP^{ms} . A total of 4200 images from 100 subjects performing each of the 6 expressions at 7 angles – from 0° to 90° in 15° intervals – were used. For this database, face segmentation was first carried

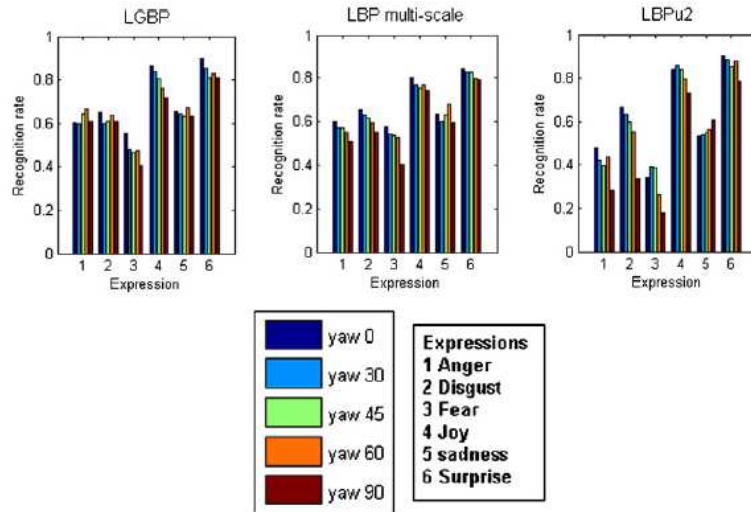


FIGURE 2.37: BU-3DFE results for the rotated view [77].



FIGURE 2.38: Six expressions in the multi-pie database [77].

out to segment the face in each image. The Viola-Jones [111] frontal face detection algorithm was used to detect frontal faces at 0° , 15° and 30° . The profile cascade was used to detect rotated faces at 45° , 60° , 75° and 90° .

Table 2.11 illustrates the average recognition accuracy results for each LBP operator at different rotation angles.

It is noted that the 15° viewing angle achieves the highest recognition results for both operators. Furthermore, as in with the previous data set, there does not appear to be any trend in the recognition accuracy with respect to the viewing angle. The recognition



FIGURE 2.39: An expression captured at different viewing angles [77].

TABLE 2.11: Multi-Pie results for various angles [77].

	0°	15°	30°	45°	60°	75°	90°
<i>LBP^{ms}</i>	76.7	80.5	70.3	69	78.6	63	73.8
<i>LGBP</i>	82.1	87.3	75.6	77.8	85	71	75.9

accuracy appears to be angle-specific. Tables 2.12 and 2.14 are confusion matrices for each of the two LBP operators.

TABLE 2.12: Confusion matrix for facial expressions over all angles for *LBP^{ms}* features [77].

(%)	Neutral	Smile	Surprise	Squint	Disgust	Scream
Neutral	73.92	11.57	2.98	8.91	3.41	0.66
Smile	9.21	78.04	4.04	4.79	3.62	1.74
Surprise	3.41	3.40	81.01	2.54	1.89	9.21
Squint	9.28	8.84	2.90	60.11	18.71	1.60
Disgust	5.51	4.85	1.74	14.87	69.21	5.27
Scream	0.15	1.15	12.95	0.94	3.48	81.57

Shan *et al.* [97] evaluated frontal FER using LBPs. The Viola-Jones [111] face detection algorithm is used to detect and segment faces. A variant of the original LBP operator

TABLE 2.13: Confusion matrix for facial expressions over all angles for *LGBP* features [77].

(%)	Neutral	Smile	Surprise	Squint	Disgust	Scream
Neutral	80.55	8.02	2.75	6.87	2.67	0.58
Smile	7.54	82.74	2.61	5.07	2.62	0.87
Surprise	1.03	3.55	88.67	0.87	1.81	5.52
Squint	8.61	7.45	1.37	66.26	16.89	0.87
Disgust	4.12	3.55	1.02	14.70	74.81	3.25
Scream	0.14	0.94	8.52	0.36	2.18	88

that uses a neighbourhood radius of 2 pixels using uniform binary patterns was used. The 110×150 pixel input images are divided into local regions using a 6×7 grid. Similar to previous researchers, a spatially enhanced LBP histogram is created by concatenating the histograms of all local regions resulting in a feature vector length of 2478.

SVMs were used for classification. Images from the Cohn-Kanade database were used in experimentation. For each subject, the neutral expression and three peak frames for each of the prototypic expressions were used in the experimentation. A total of 1280 images (108 for anger, 120 for disgust, 99 for fear, 282 for joy, 126 for sadness, 225 for surprise and 320 for neutral) from 96 subjects were used.

The dataset was partitioned into 10 equal groups, with nine groups used for training and one group used for testing. Three prominent SVM kernels were compared. Additionally, the recognition accuracy using the six prototypic expressions including the neutral expression was compared with using the six expressions without the neutral expression. The results of this comparison are summarized in Table 2.14.

TABLE 2.14: FER accuracy results of Shan *et al.* [97].

	6-Class LBP (%)	7-Class LBP (%)
SVM (linear)	91.5 ± 3.1	88.1 ± 3.8
SVM (polynomial)	91.5 ± 3.1	88.1 ± 3.8
SVM (RBF)	92.6 ± 2.9	88.9 ± 3.5

The results indicate that all three kernels can perform at a high accuracy, but the RBF kernel appears to perform marginally better in this instance. It is also noted that using both the 6-class and 7-class datasets result in very high recognition accuracies. Using a 6-class dataset appears to perform marginally better than using the 7-class dataset. However, this difference may not be statistically significant.

2.3.2 Gabor Wavelets

Gabor wavelets exhibit desirable characteristics of spatial frequency, spatial localization and orientation selectivity for image analysis [67]. Gabor wavelets can be defined as follows:

$$\psi_{\mu,\nu}(z) = \frac{\|k_{\mu,\nu}\|^2}{\sigma^2} e^{-\frac{\|k_{\mu,\nu}\|^2 \|z\|^2}{2\sigma^2}} [e - e^{-\frac{\sigma^2}{2}}] \quad (2.5)$$

where μ and ν define the scale and orientation of the Gabor kernels, $z = (x, y)$, $k_{\mu,\nu}$ is the wave vector and $\|k_{\mu,\nu}\|$ denotes the norm operator applied to $k_{\mu,\nu}$. $k_{\mu,\nu}$ is defined as follows:

$$k_{\mu,\nu} = k_\nu e^{i\phi_\mu} \quad (2.6)$$

where $k_\nu = \frac{k_{max}}{f^\nu}$ and $\sigma_\mu = \frac{\pi\mu}{8}$. The maximum frequency is k_{max} and f is the spacing factor between kernels in the frequency domain [63].

Since the Gabor kernels in Equation (2.5) can be generated from one filter – the mother wavelet – by scaling and rotation via the wave vector $k_{\mu,\nu}$, they are known as self-similar. Every kernel is a product of a Gaussian envelope as well as a complex plane wave, while the term e in square brackets determines the oscillatory part of the kernel. When the parameter σ , which determines the ratio of the Gaussian window width to wavelength, has sufficiently large values, the effect of $e^{-\frac{\sigma^2}{2}}$ in square brackets becomes negligible.

In most cases five different scales $\nu \in \{0, \dots, 4\}$ and eight orientations $\mu \in \{0, \dots, 7\}$ of Gabor wavelets are used. The real component and magnitude of the five scales and orientations of the Gabor filters with parameters $\sigma = 2\pi$, $k_{max} = \frac{\pi}{2}$ and $f = \sqrt{2}$ are illustrated in Figure 2.40.

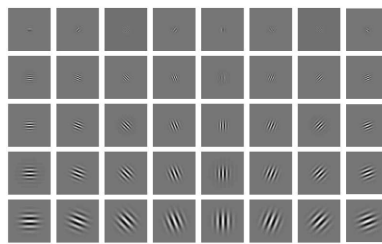


FIGURE 2.40: Gabor kernels [27].

Figure 2.41 and Figure 2.42 illustrate the resulting real component and magnitude of the Gabor representation when the kernels in Figure 2.40 are convolved with facial images. It can be seen that these representations are powerful descriptors of facial features.

The following studies have successfully implemented Gabor wavelets.

Kotsia *et al.* [61] analyzed the effects of partial occlusions of the face on frontal FER using Gabor wavelets. To recognize facial expressions, two approaches were followed:

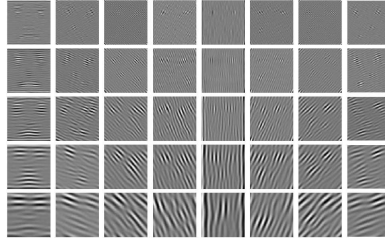


FIGURE 2.41: Real component of the Gabor representation of a facial image [27].

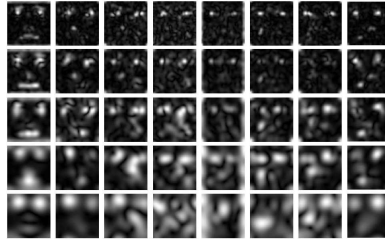


FIGURE 2.42: Magnitude of the Gabor representation of a facial image [27].

the first approach uses holistic texture information obtained by applying Gabor filters and the Discriminant Non-negative Matrix Factorization (DNMF) to the entire facial image and the second approach uses the displacement of certain points on the face using a multi-class SVM method. The Viola-Jones face detector is used to segment faces in images.

The flow diagram of the overall approach is illustrated in Figure 2.43.

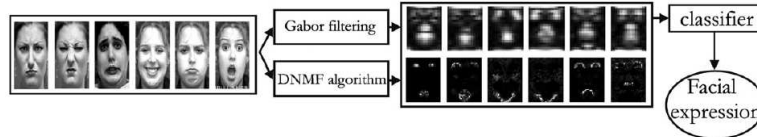


FIGURE 2.43: System overview [61].

To avoid manually selecting specific regions, Gabor filters were applied to the entire face for facial feature extraction. Four orientations were used, namely, 0 , $\frac{\pi}{4}$, $\frac{\pi}{2}$ and $\frac{3\pi}{4}$, as well as two different frequency ranges, namely, high frequencies 0, 1 and 2, and low frequencies 2, 3 and 4. A feature vector is constructed by the convolution of a 80×60 facial image with 12 Gabor filters corresponding to the orientations and frequencies mentioned. The image is down sampled to 20×15 pixels which results in a feature vector dimension of 300×1 for each of the 12 Gabor filters. In this representation, only the magnitude of the Gabor filter output was used. A combined feature vector is constructed by concatenating the 12 Gabor filter output vectors which results in a feature vector dimension of 3600×1 . The DNMF algorithm approximates a facial expression image by a linear combination of a set of basis images. Figure 2.44 depicts examples of basis images extracted for the DNMF algorithm. Since DNMF is outside the scope

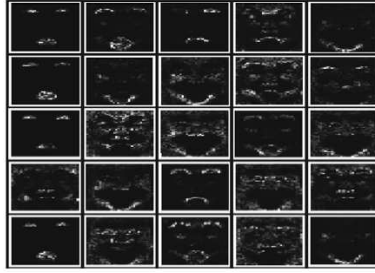


FIGURE 2.44: Basis images [61].

of this research, the reader is referred to [119] for a more detailed description of the algorithm. For the shape-based approach, a grid is placed on the face and the geometric displacement of each grid node is contained in a grid deformation feature vector. Multi-class SVMs were used for classification. Examples of the facial grid are illustrated in Figure 2.45. Two databases were used for experimentation, the image sequences from



FIGURE 2.45: Grids for each expression [61].

the Cohn-Kanade database [57] and the static images from the Japanese Female Facial Expression (JAFFE) database [73]. All of the data of both data sets was used. As explained earlier, both databases contain images on simple backgrounds.

Two experiments were carried out. The first aimed to determine the frontal FER accuracy of each method. The second experiment aimed to determine the effect of various types of occlusion on the recognition accuracy of each method. The three types of occlusions that were considered are illustrated in Figure 2.46.



(a) Eyes occluded. (b) Mouth occluded. (c) Right side of the face occluded.

FIGURE 2.46: Partial occlusions of the face [61].

Black patches were manually superimposed over the eyes, mouth and right side of the face. 80% of the data from both databases was used as training data and the rest was used as testing data. The results are summarized in Table 2.15.

TABLE 2.15: Occlusion results [61].

	No occlusion	Eyes occlusion	Mouth occlusion
JAFFE database			
Gabor filters (%)	88.1	83.1	81.5
Cohn-Kanade database			
Gabor filters (%)	91.6	86.8	84.4

The following points are to be noted from the results. As expected, the facial images that are not occluded registered the highest accuracy. The frontal results obtained from the Cohn-Kanade database registered the highest accuracy of 91.6% using Gabor filters. No results for occlusion of the right side of the face are provided. The results indicate that occlusion of the mouth appears to consistently affect the recognition accuracy of the system more than occlusion of the eyes using both databases and all methods. It is also noted that the Gabor filter method outperforms the DNMF method.

Liu and Wang [68] implemented a Gabor feature-based FER strategy to accomplish subject-independent FER of the six prototypic expressions as well as the neutral expression. Multiple orientation factors 0, 1, 2, ..., 7 and multiple scale factors 0, 1, 2, ..., 4 were chosen, which resulted in a total of 40 different Gabor filters, each with a dimension of 5×8 . The Gabor filters along with their respective response images are illustrated in Figure 2.47.

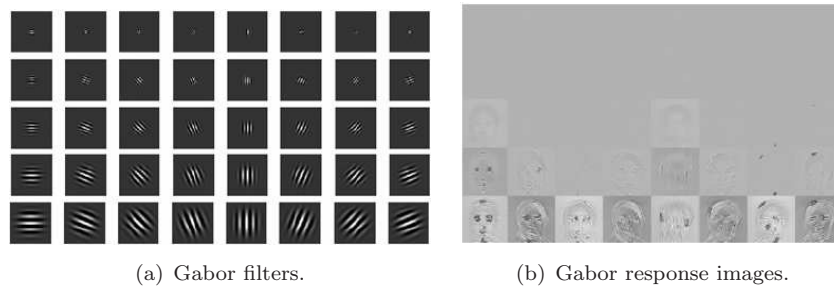


FIGURE 2.47: Gabor filters used by Liu and Wang [68].

Gabor filters which have the same scales and orientations are grouped into 13 channels corresponding to the five scales and eight orientations. Each individual channel has a unique contribution to the recognition of facial expressions. For a given test image, Principle Component Analysis (PCA) and a Neural Network are used to recognize the facial expression of each of the 13 Gabor channel-feature vectors. Two images of each expression for 10 subjects were used as training examples from the JAFFE database, the rest of the images were used for testing. Key facial landmarks were manually marked

prior to training and testing. A NN input layer consisting of 13 nodes corresponding to the 13 Gabor channel-features followed by a hidden layer of 10 TAN-SIG neurons are used. The output layer, used for classification, consists of 7 nodes. The experimental results are summarized in Table 2.16.

TABLE 2.16: Gabor filter results [68].

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Gabor PCA (%)	60	67	80	56	67	56	89

The results indicate an average accuracy of 79% across all expressions. The range in accuracy across the different expressions is (56, 89)% which indicates a large variance in accuracy.

2.4 Summary

This chapter investigated the three most prominent methods used in whole FER. These are motion-based, model-based and appearance-based methods.

Motion-based methods use motion to characterize facial expressions. Motion-based methods were divided into feature point and dense flow tracking techniques. In general, both techniques were shown to perform well with feature point tracking techniques performing better than dense flow tracking techniques. This could be due to the fact that the face is more accurately segmented in the image in feature point tracking techniques by manually placing tracking points on key facial landmarks. This is in contrast to dense flow tracking techniques which track points on an automatically generated grid overlaid on the face.

However, the reliability of motion-based methods depends mainly on the brightness constancy of pixels. Therefore, stable lighting conditions are required to achieve good results. Stable lighting conditions can not be guaranteed in real-world situations. Furthermore, these methods necessarily require image sequences of the neutral expression to a non-neutral expression to perform classification. They are not able to recognize on static images.

Model-based methods use modelling techniques to model and recognize facial expressions. Model-based methods were divided into active appearance and active shape models. In general, both methods have comparable performance in terms of accuracy. Both methods require key points on the face to be manually specified prior to modelling.

This is significant drawback of the methods since labelling can be laborious and time-consuming. They are more complex than both motion-based and appearance-based methods, require more training data, and higher computational cost is involved. In terms of accuracy, both motion-based and appearance-based methods perform better than model-based methods.

Appearance-based methods use the texture of facial images to recognize facial expressions. These methods were divided into Local Binary Patterns and Gabor filters. Both techniques are similar since they both analyze the texture at a micro-level by computing the texture descriptor in a small neighbourhood or by performing pixel-wise filtering.

In terms of accuracy, the two techniques are comparable. They are also both robust to illumination changes and variations. A great disadvantage of Gabor wavelets is the increased complexity and computational requirements as the number of kernels used increases. A large number of kernels are required to achieve good results. LBPs are not complex and are generally efficient.

Generally, appearance-based methods out-perform model-based methods but are on-par with motion-based methods. However, motion-based methods require illumination-normalized images and can only work on image sequences, whereas appearance-based methods are robust to illumination changes and can work on static images.

Therefore appearance-based methods and specifically Local Binary Patterns are selected as the feature extraction method for this research.

2.5 Conclusion

It is concluded that the LBP appearance-based method is the most suitable facial feature extraction method, and is selected for use in this research.

Of special note were the studies by Moore and Bowden [77] and Kotsia *et al.* [61]. Moore and Bowden optimized the resolution of facial images before applying the LBP operator towards FER. They also investigated the effect of various yaw rotations of the face on the FER accuracy. Both of these experiments are adapted for use in this research. Furthermore, it was noted that there is no trend in the recognition accuracy with respect to the viewing angle. The recognition accuracy appears to be angle-specific. Thus, the system needs to be trained and tested on a per-angle basis. For the scope of this research, the frontal view and one rotated angle of 60° are therefore selected.

Furthermore, it was found that of the purely LBP methods, the multi-scale LBP^{ms} and uniform LBP^{u2} LBP operators perform the best. A hybrid between these two operators is used in the proposed FER strategy.

It was also noted that the accuracy of their strategy varied between the BU-3DFE database and the Multi-pie database. This shows that the accuracy is affected by the nature of the data. This research uses the BU-3DFE database in training and testing.

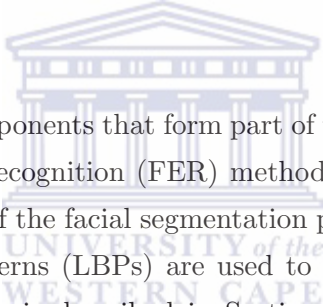
Kotsia *et al.* compared the effects of different types of occlusion on FER accuracy using Gabor filters. This research uses an extended version of this experimentation to investigate the effects of various types and levels of occlusion on FER accuracy using LBPs.

Finally, it was noted that the majority of research across all three methods makes use of manually segmented images. Few automatic segmentation strategies are proposed, hence the need to lay the foundation towards a fully automatic segmentation approach.



Chapter 3

Image Processing in Appearance-Based Facial Expression Recognition



This chapter discusses the components that form part of the appearance-based approach used in the facial expression recognition (FER) methodology of this research. Section 3.1 discusses the components of the facial segmentation process. Once the face has been segmented, Local Binary Patterns (LBPs) are used to extract features from the face. Feature extraction using LBPs is described in Section 3.2. Using the facial features extracted, an SVM is used to carry out facial expression classification. In Section 3.3, a detailed discussion on SVMs is provided. The chapter is then concluded.

3.1 Face Segmentation Techniques

3.1.1 Face Detection

Face detection is a popular initial step in FER systems. It is used to localize and extract the face region in an image from the background [45]. The Viola-Jones [111] object detection framework is a popular and robust implementation of face detection. This tree-based approach uses Haar feature classifiers to build a boosted rejection cascade. At every node in the cascade, AdaBoost is used to achieve a high detection rate. This method consists of four novel features [12]:

1. The use of Haar-like wavelet features to characterize the face.

2. An intermediate image representation, referred to as an integral image, for the rapid computation of Haar-like wavelet features.
3. A learning algorithm, based on AdaBoost, which yields extremely efficient classifiers.
4. A rejection cascade consisting of a combination of weak classifiers, which focus on object-like regions and discard background regions of the image.

These four features are discussed in the following subsections.

3.1.1.1 Haar-like Wavelets

Haar-like wavelets or Haar-like features are single wavelength square waves – one high and one low interval. Figure 3.1 illustrates three different types of Haar-like features.

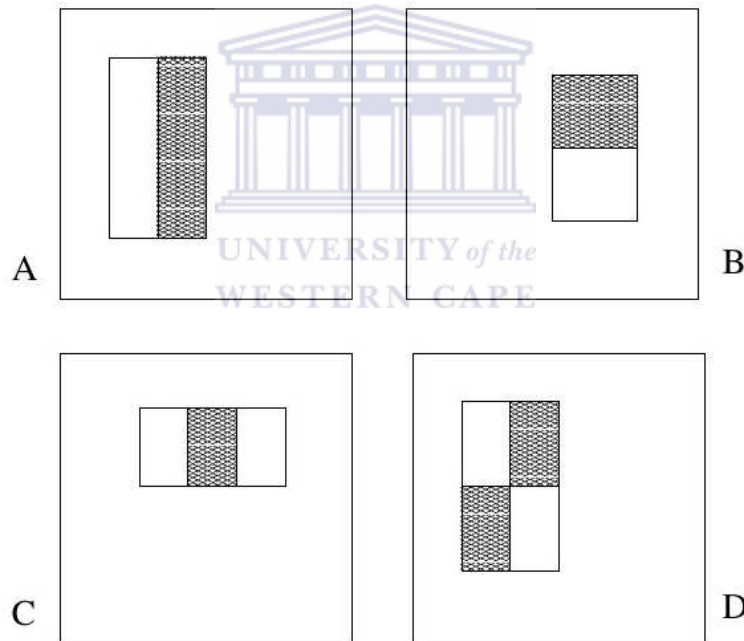


FIGURE 3.1: Haar-like features.

Two-rectangle features, depicted in block A and block B, consist of rectangles which have identical size and shape, and are horizontally or vertically adjacent. These features are scanned over an image at different scales and regions. For the two-rectangle feature, the difference between the sum of the pixels in the image within the two rectangular regions is computed. For three-rectangle features, depicted in block C, the sum of the image pixels within the two outside rectangles is subtracted from the sum of the image pixels in the centre rectangle. For four-rectangle features, depicted in block D, the difference between the sum of the pixels in the diagonal pairs of rectangles is computed [112].

3.1.1.2 Integral Image

Scanning the Haar-like features over an image and computing the relevant pixel sums at different scales can be very computationally expensive. The use of larger Haar-like features results in higher computational overhead than the use of smaller scales.

An integral image is a data structure used to efficiently determine the presence or absence of numerous Haar-like features at every image location and at several scales in constant time. An integral image is computed by taking the sum of all the pixels above and to the left of a corresponding pixel. Consider an input image I of dimension $W \times H$. The resulting integral image I' then has a dimension of $(W + 1) \times (H + 1)$. A buffer of zero values along the x -axis and y -axis are inserted, which is required for efficient computation[12].

Starting at the top-left pixel in Figure 3.2(a), the integral pixel values in Figure 3.2(b) are calculated using the following formula [12]:

$$I'(x, y) = I(x, y) + I'(x - 1, y) + I'(x, y - 1) - I'(x - 1, y - 1) \quad (3.1)$$

An example configuration of pixel values in the original image is illustrated in Figure 3.2(a) and the corresponding integral image is illustrated in Figure 3.2(b).

1	2	5	1	2	1	3	8	9	11
2	20	50	20	5	3	25	80	101	108
5	50	100	50	2	8	80	235	306	315
2	20	50	20	1	10	102	307	398	408
1	5	25	1	2	11	108	338	430	442
5	2	25	2	5	16	115	370	464	481
2	1	5	2	1	18	118	378	474	492

(a) Pixels values in the original image.

(b) Computed values in the integral image.

FIGURE 3.2

Using the integral image, the computation of any Haar-like feature at any location and scale can be achieved in constant time. For example, in Figure 3.3, the sum of pixels in any arbitrary rectangular region D can be computed by subtracting the integral image values at points 2 and 3 from the value at point 4, and adding the value at point 1. This can be extended to compute Haar-like features. Any two-rectangle feature can be computed using 8 references to the integral image, and 9 references are required to compute four-rectangle features.

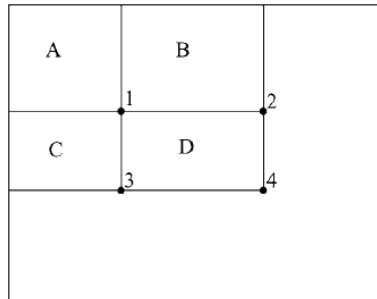


FIGURE 3.3: Optimization of the Haar-like feature using the integral image.

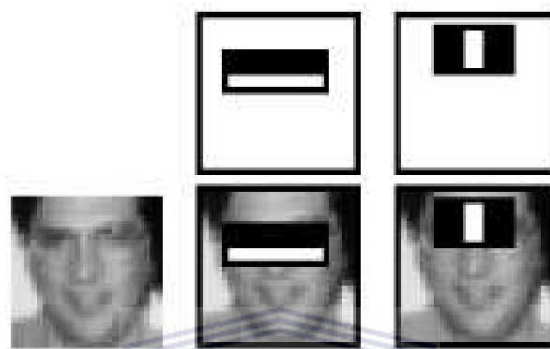


FIGURE 3.4: AdaBoost feature selection.

3.1.1.3 AdaBoost Learning Algorithm

Viola and Jones [111, 112] used a modified AdaBoost algorithm to select a small set of features as well as train a classifier. The AdaBoost algorithm, in its original form, is used to boost the classification performance of a simple or weak learning algorithm. A strong classifier is created by combining many weak classifiers. This process is known as boosting and involves assigns weights to each weak classifier, with the best weak classifier selected at each boosting interval. This results in a strong classifier which consists of a combination of weighted classifiers. Figure 3.4 illustrates an example of features selected by the AdaBoost algorithm.

The two features in the top row are the first and second features selected by AdaBoost. The first feature measures the difference in intensity between the region of the eyes and upper cheeks. The second feature measures the difference in intensity between the eyes and the bridge of the nose.

3.1.1.4 Constructing a Cascade of Weak Classifiers

The Viola-Jones face detection algorithm [111] constructs smaller and more efficient boosted classifiers to reject many negative sub-windows while detecting many positive

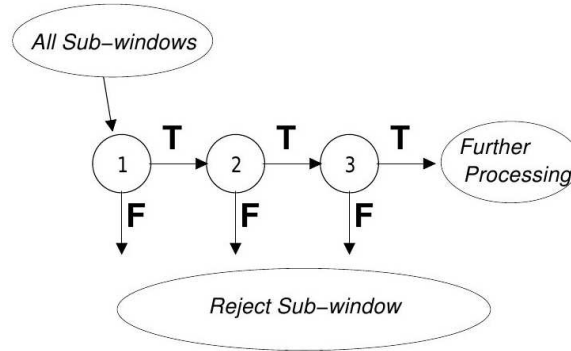


FIGURE 3.5: Rejecting regions in an image.

instances. Initially, weak classifiers are used to reject most of the sub-windows prior to using more complex classifiers to achieve low false positive rates. Figure 3.5 illustrates the detection process taking the form of a degenerate tree, which is also known as a cascade. Starting at the first classifier, a positive result triggers the classification of a second classifier. A positive result from the second classifier triggers the classification of a third classifier, and so on. At any stage, a negative result triggers the immediate rejection of a sub-window.

Classifiers are trained for each stage using AdaBoost. The default AdaBoost threshold is designed to yield a low error rate. However, in order to minimize false negatives even further, the boosted classifiers are adjusted. Higher detection and false positive rates are thus obtained using a lower threshold.

3.1.1.5 Testing the Viola-Jones Face Detection Algorithm

Viola and Jones tested the algorithm using 507 labelled frontal images from the MIT-CMU frontal face test set [111]. The images contain subjects with varied skin tones on various complex backgrounds. Figure 3.6 depicts the results of applying the Viola-Jones face detection algorithm on the images. The algorithm obtained an accuracy of 93.9% at a real-time speed of 15 FPS on a Pentium III 700 MHz PC. This result is highly encouraging since the face detection procedure forms the basis of the face segmentation procedure in this research. Therefore, this face detection technique will be employed in this research for the detection of frontal faces.

3.1.2 Morphological Operations

A morphological operator usually takes a binary image and a structuring element as input. The input is combined and used as a set operator, i.e. intersection, union, compliment etc. [35].

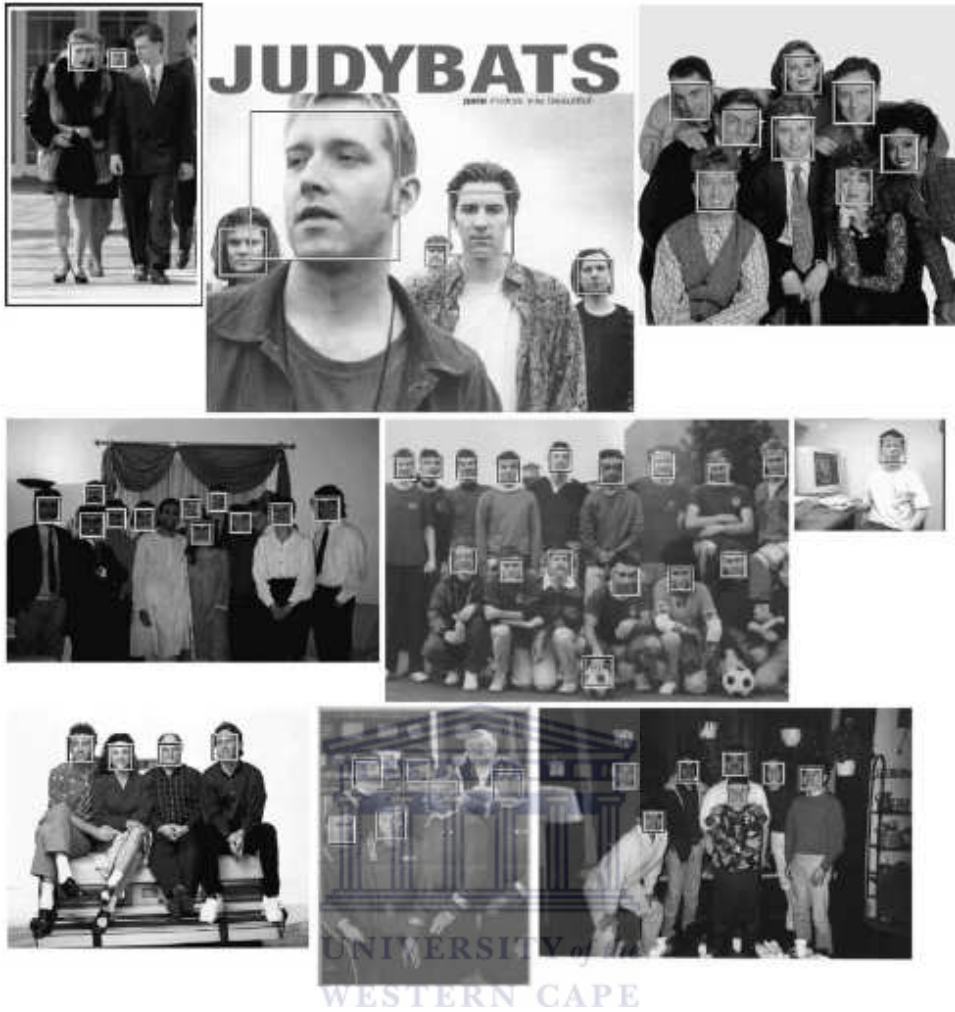


FIGURE 3.6: Positive faces detected by the Viola-Jones algorithm [111].

The structuring element encodes characteristics of the shape of an object within an image. It consists of a pattern which specifies the coordinates of a number of distinct points relative to a particular origin. Since cartesian coordinates are usually used, the element can conveniently be represented by a small image on a rectangular grid. An example of a number of different structuring elements of different sizes are depicted in Figure 3.7.

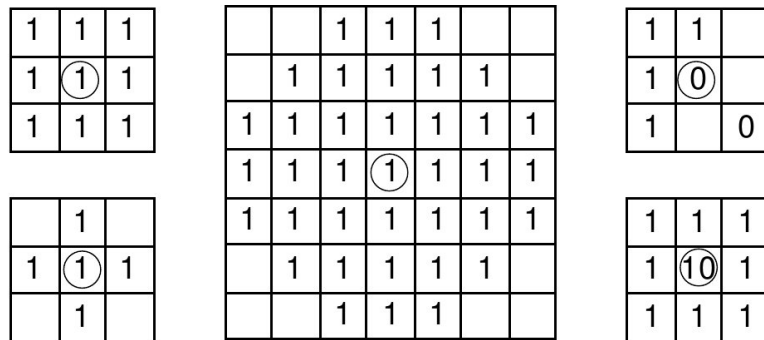


FIGURE 3.7: Examples of various structuring elements [35].

In each case, a ring around a point emphasizes the origin. As illustrated by the figure, structuring elements that fit into a 3×3 grid are the most commonly seen types. Note that, although a rectangular grid is used to represent a structuring element, not every point in the grid is necessarily part of the structuring element. Thus, some elements contain blanks.

In the application of a morphological operation, the origin of the structuring element is translated to each pixel location in the image. The points contained in the translated structuring element are then compared with the pixel values in the underlying image.

The following subsections discuss the dilation and erosion operations.

3.1.2.1 Dilation

Applying this operator to a binary image results in the enlargement of the boundaries of foreground pixels, typically white pixels. Therefore, regions of foreground pixels grow in size while the holes within them become relatively smaller.

The symbol \oplus represents the dilation operator, which is applied to a binary image I by a structuring element B , which can be mathematically formulated as:

$$I \oplus B = \bigcup_{b \in B} I_b = \{x \mid (B^s)_x \cap I \neq \emptyset\} \quad (3.2)$$

where B^s denotes the reflection of the set B and $(B^s)_x$ is B^s translated by the vector x .

3.1.2.2 Erosion

Applying this operator to a binary image results in the shrinking of boundary regions of foreground pixels. Therefore, regions of the foreground pixels shrink in size while the holes within them become relatively larger.

The symbol \ominus represents the erosion operator, which is applied to a binary image I by a structuring element B , which can be mathematically formulated as:

$$I \ominus B = \{x \mid (B)_x \subseteq I\} \quad (3.3)$$

where $(B)_x$ is the set B translated by the vector x .

3.1.3 Eye Detection

Eye detection is an essential step in many applications such as face recognition, facial expression analysis, eye-gaze estimation, criminal investigation, human interactions and surveillance systems [21, 28, 41]. Eye detection methods can generally be classified into three categories: template-based methods [32, 51], appearance-based methods [50, 90] and feature-based methods [1, 59]. The question as to which eye detection method is the best is unclear from the literature. However, most researchers [26, 38, 49, 83, 84, 121] use the template-based approach to determine the exact location of the eyes accurately. This method is explained in this section.

The template-based approach constructs eye maps from a facial image. The original image is transformed from RGB to the YC_bC_r colour space [83]. The details of these colour spaces are explained in the a subsequent section. Initially, two separate eye maps are constructed from the facial image, $EyeMapC$ from the two chrominance components and $EyeMapL$ from the luminance component. These eye maps are combined into a single eye map called $EyeMap$.

3.1.3.1 EyeMapC

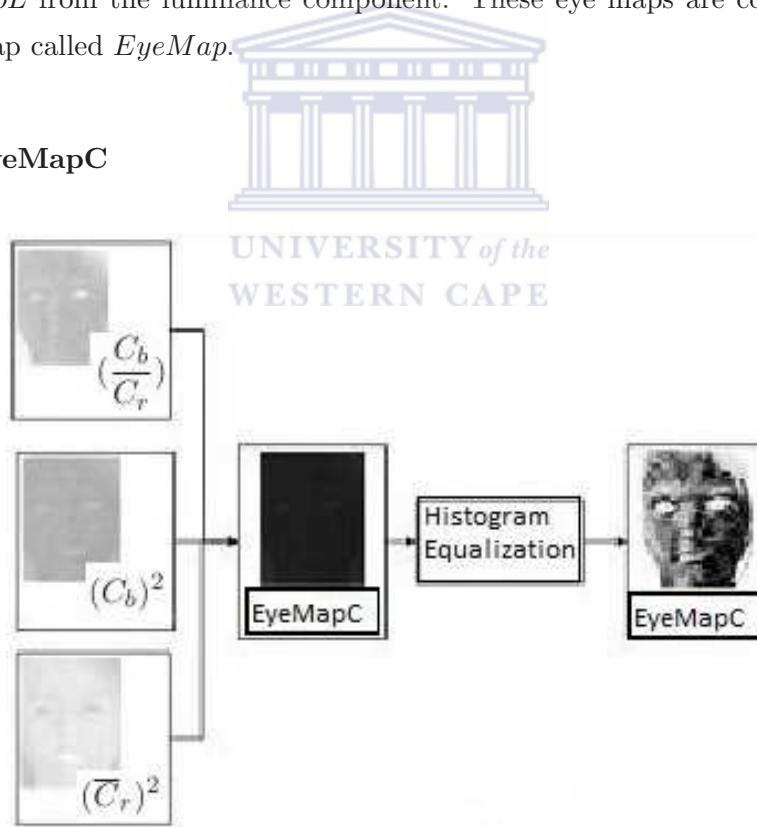


FIGURE 3.8: The construction of $EyeMapC$.

The chrominance channels C_b and C_r of the YC_bC_r colour space contain high C_b and low C_r values in the eye regions [49]. This eye map is constructed by applying the following formula to each pixel (x, y) in the facial image:

$$EyeMapC = \frac{1}{3} \left\{ (C_b)^2 + (\overline{C_r})^2 + \left(\frac{C_b}{C_r} \right) \right\} \quad (3.4)$$

where $(C_b)^2$, $(\overline{C_r})^2$ and $(\frac{C_b}{C_r})$ are normalized to the range $[0, 1]$ and $\overline{C_r}$ is the additive inverse of C_r . This operation highlights pixels with high C_b and low C_r values. Pixels with higher C_b values are emphasized by the term $(C_b)^2$ which also causes pixels with lower C_b values to become less pronounced.

Furthermore, $(\frac{C_b}{C_r})$ causes low C_r values to become brighter. The scaling factor $\frac{1}{3}$ is applied to ensure that the resulting eye map $EyeMapC$ remains in the range $[0, 1]$. Finally, histogram equalization is carried out on $EyeMapC$, illustrated in Figure 3.8.

3.1.3.2 EyeMapL

The two morphological operators, dilation and erosion, are applied to emphasize brighter and darker pixels in the luminance component Y of the YC_rC_b colour space [53]. $EyeMapL$ is constructed by applying the following formula:

$$EyeMapL = \frac{Y \oplus B}{Y \ominus B} \quad (3.5)$$

where \oplus and \ominus are the dilation and erosion operations explained in the previous section, and B is the structuring element.

3.1.3.3 Eye Map

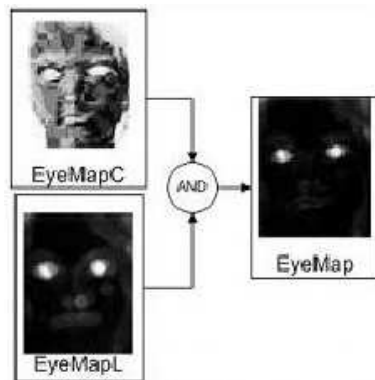


FIGURE 3.9: The construction of the final eye map.

The final eye map, $EyeMap$, illustrated in Figure 3.9 is constructed by applying the following formula:

$$EyeMap = (EyeMapC) \cap (EyeMapL) \quad (3.6)$$

The resulting image consists of only a pair of illuminated eyes.

3.1.4 Skin Detection

Skin colour has proven to be a useful and robust cue for face detection, localization and human tracking [3, 13, 65]. Since most skin tones are distinct from the colours of most objects, specific body parts can be tracked using this information [82]. Skin colour information can be considered a very effective tool for the identification of facial areas provided that the underlying skin-colour pixels can be represented, modelled and classified accurately.

Skin detection classifies each individual pixel in the image as being either a skin pixel or a non-skin pixel [13]. Several factors such as illumination, camera properties and the viewing angle make skin detection a non-trivial process. The three primary steps for skin detection using colour information are [56]:

1. Selecting a suitable colour space for the representation of image pixels.
2. Selecting a suitable distribution for modelling skin and non-skin pixels.
3. Classifying pixels in the image as either skin or non-skin pixels.

Computer graphics and video signal transmission standards have given birth to a wide variety of colour spaces with different properties [109]. The following subsections discuss the most popular colour spaces in order to determine and justify the use of a colour space conducive to the task of skin detection.

3.1.4.1 RGB Colour Space

RGB is the acronym for Red-Green-Blue and it is a colour space which originated from cathode ray tubes (CRT) and display graphics when colour was described as combinations of these three coloured rays [109]. The colour of a single pixel can be represented by this combination and the channels are highly correlated. The luminance and chrominance data is therefore not separated. Other colour spaces are obtained by performing a linear or non-linear transformation on the RGB colour space. Applying a colour space transformation can help reduce the overlap between the luminance and chrominance information. This can contribute towards robustness to varying illumination conditions.

3.1.4.2 Normalized RGB Colour Space

The normalized RGB colour space can be obtained by applying the following transformation to the default RGB colour space:

$$r = \frac{R}{R + G + B}, \quad g = \frac{G}{R + G + B}, \quad b = \frac{B}{R + G + B} \quad (3.7)$$

where r, g and b are the normalized red, green and blue pixels and R, G and B are the red, green and blue pixel values from the RGB colour space. Note that the sum of the normalized pixel values is 1:

$$r + g + b = 1 \quad (3.8)$$

The sum is constant, hence the third component – b – can be omitted as it does not hold any significant information. The space dimensionality is effectively reduced by this omission.

3.1.4.3 HSV Colour Space

HSV is the acronym for Hue-Saturation-Value and is also known as HSI (Hue-Saturation-Intensity) and HSL (Hue-Saturation-Lightness). This colour space describes colour based on the artist's idea of tint, saturation and tone [109]. Hue defines the dominant colour of a region and saturation measures the colourfulness of a region in proportion to its brightness. The intensity, lightness or value is related to luminance. The explicit discrimination between luminance and chrominance makes this colour space a popular choice for skin colour segmentation [8, 55, 75, 101, 120]. The mapping of the RGB colour space to the HSV colour space is achieved by a non-linear transformation which is formulated as follows [115]:

$$V = \max(r, g, b) \quad (3.9a)$$

$$S = \frac{\max(r, g, b) - \min(r, g, b)}{V} \quad (3.9b)$$

$$H = \begin{cases} \frac{g - b}{6(\max(r, g, b) - \min(r, g, b))}, & \text{if } V = r \\ \frac{2 - r + b}{6(\max(r, g, b) - \min(r, g, b))}, & \text{if } V = g \\ \frac{4 - g + r}{6(\max(r, g, b) - \min(r, g, b))}, & \text{if } V = b \end{cases} \quad (3.9c)$$

where H, S and V are the Hue, Saturation and Value components, r, g and b are the normalized red, green and blue pixel values, and $\max(r, g, b)$ and $\min(r, g, b)$ are the maximum and minimum between the normalized red, green and blue pixel values. The Hue component is illumination invariant which makes it less sensitive to lighting changes than the other components [6].

3.1.4.4 YC_bC_r Colour Space

The YC_bC_r colour space is an encoded non-linear RGB signal and is often used in European television networks [3]. Its colour is represented by luminance, constructed as a weighted sum of the RGB values and two colour difference values C_r and C_b that are formed by subtracting the luminance component from the RGB red and blue components. This can be formulated as:

$$Y = 0.299R + 0.587G + 0.114B \quad (3.10a)$$

$$C_r = R - Y \quad (3.10b)$$

$$C_b = B - Y \quad (3.10c)$$

where Y represents the luminance component and C_r and C_b represent the chrominance components, respectively. This colour space is also suitable for skin detection since skin colours of different races are found to occur in the chrominance channels [31]. It is possible to discard the Y component for skin detection, since the luminance component is easily separable from the chrominance components.

3.1.4.5 TSL Colour Space

TSL is the acronym for Tint-Saturation-Lightness which is a chrominance-luminance colour space and a transformation of the normalized RGB colour space. The TSL colour space can be formulated as follows:

$$T = \begin{cases} \frac{\arctan(\frac{r'}{g'})}{2\pi} + \frac{1}{4}, & \text{if } g' > 0 \\ \frac{\arctan(\frac{r'}{g'})}{2\pi} + \frac{3}{4}, & \text{if } g' < 0 \\ 0, & \text{if } g' = 0 \end{cases} \quad (3.11a)$$

$$S = \sqrt{\frac{9(r'^2 + g'^2)}{5}} \quad (3.11b)$$

$$L = 0.299R + 0.587G + 0.114B \quad (3.11c)$$

where T , S and L represent the Tint, Saturation and Lightness of a pixel, r' and g' represent variants of the normalized red and green pixel values, given by:

$$r' = r - \frac{1}{3} \quad (3.12a)$$

$$g' = g - \frac{1}{3} \quad (3.12b)$$

3.1.4.6 A Colour Space Conducive to Skin Detection?

Many researchers take two factors into consideration when deciding on a colour space transformation conducive to skin detection.

1. The colour space transformation should aid the separation of skin and non-skin pixels.
2. It should be illumination invariant – address the problem of varying lighting conditions.

Four studies have specifically been carried out to investigate the effectiveness of colour space transformation for the purpose of skin detection and to ascertain whether the aforementioned assumptions hold [56, 100, 109, 120].

Kakumanu *et al.* [56] reviewed critical skin detection issues in their research. They concluded that non-parametric methods, such as histogram-based methods, are generally not affected by the colour space representation. Yet, parametric modelling approaches, such as Gaussian modelling, are affected by the colour space representation. These methods are better suited for constructing classifiers in the case where training data is limited. Furthermore, they concluded that there was no significant improvement to the skin detection process when using the RGB colour space as opposed to using a non-RGB colour space. Shin *et al.* concluded that the separability between the two classes of skin and non-skin pixels was highest in the RGB colour space. Their findings suggest that the separation of the luminance and chrominance components decreases the separability of skin and non-skin pixels significantly.

However, research carried out by Vezhnevets *et al.* [109] suggests that the exclusion of the luminance component only aids the generalization of sparse training data. Zarit *et al.* [120] suggested that a colour space transformation conducive to skin detection should be based on whether any form of post-processing requires a particular colour space. Although the general question as to which colour space is the most appropriate

for skin detection remains unanswered, their research concludes that the HSV colour space aids the skin detection process.

Based on the research mentioned above, it is unclear whether a non-RGB colour space is conducive to skin detection. Many researchers [23, 24, 31, 102] agree with Forsyth and Fleck [36]. Their research indicates that the colour range in the Hue component of the HSV colour space is representative of human skin colour. The colour of human skin is formed by a combination of carotene, haemoglobin and melanin. Carotene contains a peculiar yellow-orange colour which is mostly found in the palms and soles. Haemoglobin, the substance carrying oxygen in red blood cells, is responsible for the pink-red colour in the skin. Melanin is the primary determinant of skin colour. The Hue component in the HSV colour space can easily differentiate between the combination of colours found in two types of melanin: pheomelanin which is of a red colour and eumelanin which is of a very dark brown colour.

While it is beyond the scope of this research to establish which colour space is optimal for skin detection, based on the above research it is concluded that the Hue component can effectively be used to characterize skin colour amongst all races and skin tones [23, 24, 31, 102].

3.1.4.7 Skin Model



FIGURE 3.10: Original colour image.

Studies have shown that South Africa, as well as the sub-Saharan African populations, have the highest skin colour diversity in the world [94]. Achmed [3] proposed a dynamic

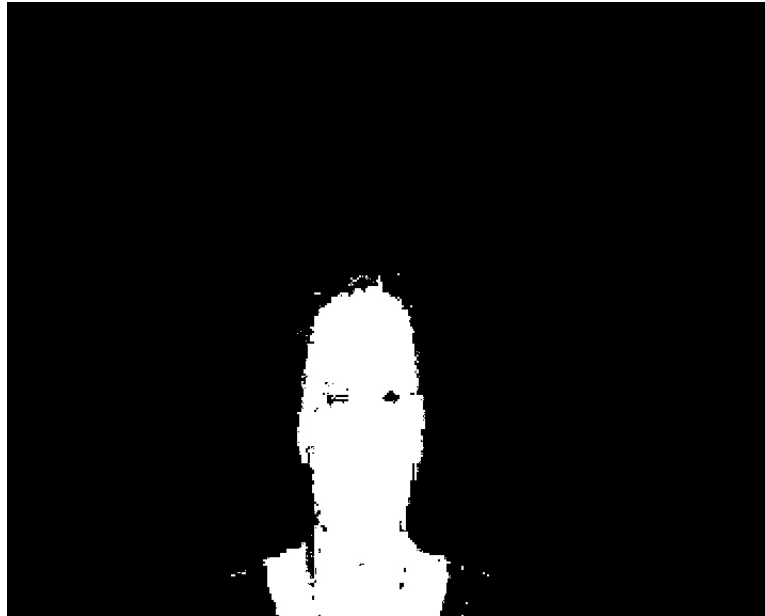


FIGURE 3.11: Skin image.

skin detection method. This method identifies skin pixels amongst all races and varied skin tones dynamically. This approach locates a 10×10 pixel region at the centre of the nose and uses this pixel distribution to represent the overall skin tone of an individual. This region is used because it is not affected by facial hair, eyes, lips or spectacles. The Hue values in this region are used to create a histogram which serves as a look-up table for the distribution of skin pixels.

The histogram groups the pixel values into a set of predefined bins [12]. The bin width corresponds to the number of data points that are assigned to each bin. Brown [13] optimized Achmed's skin detection method [3] by optimizing the bin width. Bin widths ranging from 4 to 32 in increments of 4 were compared. The skin pixel true positive detection rate deteriorated as the bin width increased. This was attributed to the fact that a larger bin width causes significant loss of detail due to grouping a larger range of pixels into a small number of bins. The research concluded that a bin width of 8 registered the highest combination of true positive and true negative skin detection accuracy. Therefore, a bin width of 8 will be adopted in the skin detection implementation of this research.

The resulting Hue histogram is back-projected onto the original image. This results in the production of a new greyscale image consisting of intensity values ranging from 0 to 255. The value 0 and 255 at each pixel location indicate, respectively, the lowest and highest likelihood that that pixel location is of skin colour within the original image. Thresholding is used to binarize this image into skin and non-skin classes. A threshold of 60 was found to yield accurate skin detection results by Achmed [3] and Li [65].

Figure 3.10 illustrates an input image. Figure 3.11 illustrates the result of applying the skin detection method to the image depicted in Figure 3.10.

3.2 Feature Extraction Using Local Binary Patterns

Local Binary Pattern (LBP) features were originally proposed for texture analysis [86, 87], but have recently been introduced to represent faces in facial image analysis [4, 33, 44]. The two most important properties of LBP features are their tolerance to illumination changes and their computational simplicity [97].

3.2.1 The Original LBP Operator

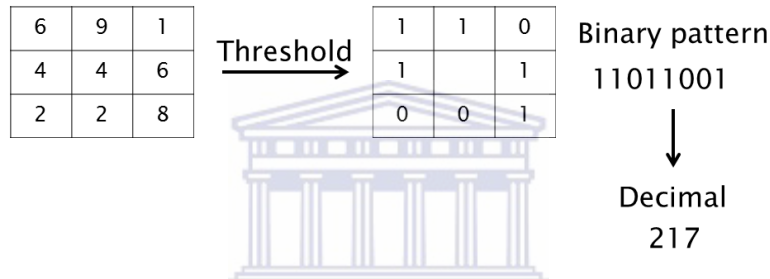


FIGURE 3.12: The original LBP operator.

The original LBP operator was described in the previous chapter but a brief description is provided here for completeness and convenience. The operator is applied to each 3×3 neighbourhood of pixels in an image. Referring to Figure 3.12, the operator, starting at the top left corner and progressing clockwise, applies a threshold to each pixel in the 3×3 neighbourhood of the centre pixel by thresholding the neighbouring pixel using the value of the centre pixel as the threshold.

The threshold function, S , is formulated mathematically as follows:

$$S(f_p, f_c) = \begin{cases} 1 & \text{if } f_p \geq f_c \\ 0 & \text{otherwise} \end{cases} \quad (3.13)$$

where $\{f_p | p = 0, \dots, 7\}$ are the neighbouring pixels and f_c is the centre pixel. Then, by assigning a binomial factor 2^p for each $S(f_p, f_c)$, the LBP value of the centring pixel is calculated with the following formula:

$$LBP(f_c) = \sum_{p=0}^7 2^p (S(f_p, f_c)) \quad (3.14)$$

These derived binary numbers codify local primitives – curved edges, spots and flat areas – so that each LBP code can be regarded as a micro-texton [97]. Examples of texture primitives are illustrated in Figure 3.13.

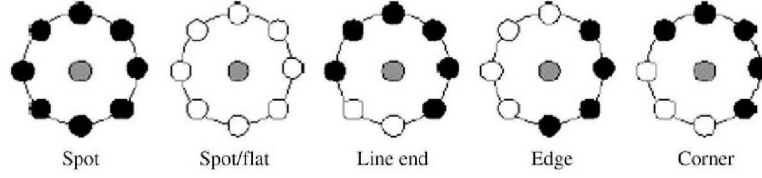


FIGURE 3.13: Examples of texture primitives [97].

3.2.2 LBP Histograms

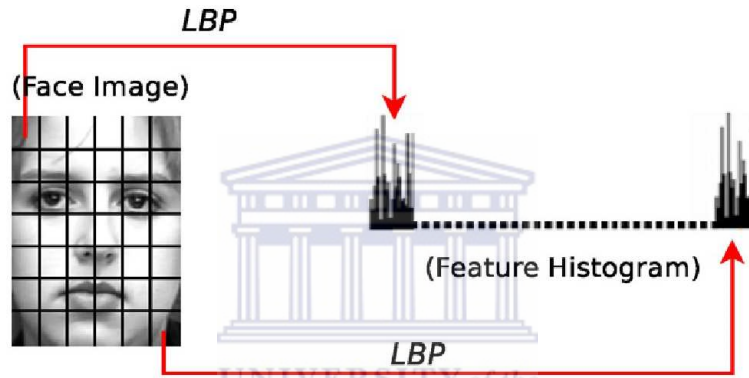


FIGURE 3.14: Concatenating region histograms into one single, spatially enhanced histogram [97].

After an image is labelled with a specific LBP operator, a histogram H of the labelled image $f_l(x, y)$ is computed using the following formulation:

$$H(i) = \sum_{x,y} I\{f_l(x, y) = i\} \quad (3.15)$$

where $i \in \{0, \dots, n - 1\}$ is a bin number, n is the number of bins in the histogram and I is formulated as:

$$I\{A\} = \begin{cases} 1, & \text{if } A \text{ is true} \\ 0, & \text{if } A \text{ is false.} \end{cases} \quad (3.16)$$

The resulting histogram contains information about the distribution of micro-patterns over the entire image. However, the retention of spatial information is required for effective facial representation. This is achieved by dividing the labelled image into m regions R_0, \dots, R_{m-1} . This spatially enhanced histogram is formulated as:

$$H(i, j) = \sum_{x, y} I\{f_i(x, y) = i\} I\{(x, y) \in R_j\} \quad (3.17)$$

where $i \in \{0, \dots, n - 1\}$ is a bin number and n is the number of bins. In this histogram, the face is effectively described in terms of three levels of locality [4]. The histogram labels contain information about the patterns on a pixel level, the labels are summed over regions to produce information on a regional level and the regional histograms are concatenated to build a global description of the face. Figure 3.14 illustrates an image divided into regions from which the LBP histograms are constructed and concatenated into a spatially enhanced histogram.

3.2.3 The Extended LBP Operator

The original LBP operator is limited by its 3×3 neighbourhood and therefore, it cannot capture dominant features with large scale structures. Hence, the operator was later extended. The extended LBP operator, also known as the multi-scale LBP operator, was proposed by Ojala *et al.* [87] to use neighbourhoods of different sizes. The use of circular neighbourhoods and the bilinear interpolation of pixel values allows for the use of any radius and any number of pixels within the neighbourhood. Figure 3.15 illustrates the extended LBP operator $LBP_{(P,R)}$ where the notation (P, R) denotes a neighbourhood of P equally spaced sampling points on a radius R which forms a circularly symmetric neighbour set.

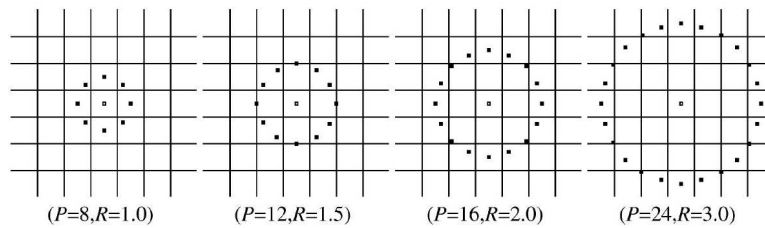


FIGURE 3.15: Examples of texture primitives [97].

Research [4, 5, 7, 44, 97, 98] has consistently shown that the $(P = 8, R = 2)$ neighbourhood results in a very accurate description of facial features for FER, therefore, the $LBP_{8,2}$ operator will be implemented in this research.

3.2.4 Uniform and Non-Uniform Patterns

A total of 2^P different binary patterns can be obtained using P pixels in the neighbourhood set. Therefore, the $LBP_{(P,R)}$ operator produces 2^P different output values.

This results in 2^P bins in the histogram of the LBP image computed. Research has shown that specific bins in the histogram contain significantly more information than others [87]. Thus, it is possible to use only a specific subset of the 2^P binary patterns to accurately describe the texture of images at an increased efficiency. These fundamental patterns are known as uniform patterns [77]. A pattern is uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa. Figure 3.16 illustrates a graphical example of a uniform and a non-uniform pattern.

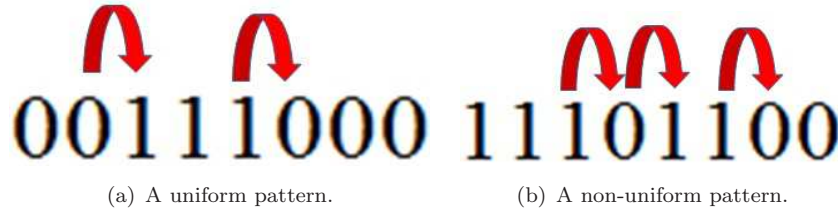


FIGURE 3.16

Figure 3.16(a) contains exactly two transitions, so it is deemed a uniform pattern. Figure 3.16(b) contains more than two transitions, so it is deemed a non-uniform pattern. Research has shown that uniform patterns account for nearly 85.2% of all patterns in the $(P = 8, R = 2)$ neighbourhood and about 70% in the $(P = 16, R = 2)$ neighbourhood in texture images [87].

The use of uniform patterns makes it possible to accumulate all non-uniform patterns in the LBP histogram into a single combined bin, while retaining separate bins for uniform patterns. This yields a uniform LBP operator denoted $LBP_{P,R}^{u2}$ which contains less than 2^P bins in the resulting histogram [97]. For example, the number of bins for a neighbourhood of 8 pixels is 256 for the standard LBP operator, but only 59 – 58 uniform bins and 1 combined non-uniform bin – for $LBP_{P,R}^{u2}$, which is considerably less. This procedure results in significantly lower computational cost and higher efficiency at a high accuracy.

3.3 Support Vector Machines

Support Vector Machines (SVMs) are known to be popular machine learning tools used especially for solving pattern recognition problems [3, 13, 65, 85, 114]. The derivation of SVMs stem from statistical learning theory. SVMs were initially intended towards binary classification problems in which data was only classified into two classes. However, they have been extended to solve multi-class problems.

SVMs offer several significant advantages when compared to other classifiers [114]. One advantage is the training time that is not affected by the use of large images which

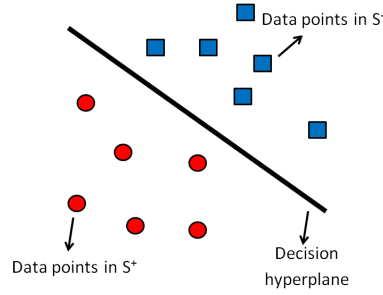


FIGURE 3.17: Decision boundary for the linear classification case.

produce feature vectors of high dimensionality. Another advantage is attributed to the use of its kernel functions which provide power and flexibility. The popular open-source SVM tool – LibSVM [17] – uses the Radial Basis Function (RBF) as its default kernel. However, other kernels such as the linear kernel, polynomial kernel, sigmoid kernel and the precomputed kernel, offer alternatives to the default kernel. The application of the aforementioned alternative kernels may result in a more even spread of the data, thus, allowing non-linear classification problems to be solved using linear classification techniques.

The sections below describe the underlying theory of SVM classification.

3.3.1 The Optimal Hyperplane

SVMs intend to maximize a mathematical function given a collection of data points [85]. It is possible to separate data points that consist of two classes by finding a boundary between these classes. Consider S as a set of M training points which is expressed as $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_M, y_M)\}$. Let $i \in \{1, 2, \dots, M\}$, therefore, each x_i is a data point in R^n and each $y_i \in \{-1, 1\}$ is the label corresponding to the data points, which is divided into separate positive and a negative classes.

Consider the positive and negative classes $S^+ = \{x_i \mid y_i = 1\}$ and $S^- = \{x_i \mid y_i = -1\}$, respectively, that are linearly separable in R^n . The training of the SVM results in at least one boundary – the decision boundary – that can be formed between the two classes [85]. The decision boundary is illustrated in Figure 3.17. The decision boundary, in higher-dimensional space, takes the form of a plane which is illustrated in Figure 3.18.

This plane is referred to as the decision hyperplane and can be formulated as follows:

$$f(x) = \mathbf{w} \cdot x + b = 0; \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R} \quad (3.18)$$

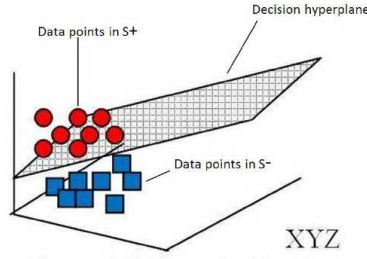


FIGURE 3.18: The decision boundary in higher-dimensional space.

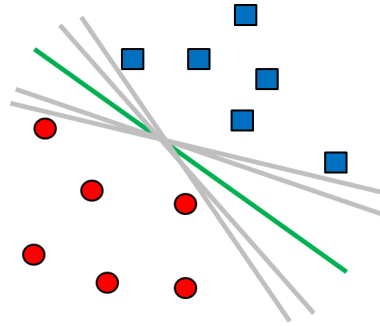


FIGURE 3.19: The decision hyperplane.

where \mathbf{w} is the normal vector and b is the interim term. Vector \mathbf{w} of the decision hyperplane is defined as a linear combination of x_i with weights α_i as follows:

$$\mathbf{w} = \sum_{i=1}^M \alpha_{\alpha_i} x_i y_i \quad (3.19)$$

Numerous decision boundaries are able to separate the two classes, as illustrated in Figure 3.19. It is important to note that the green line in Figure 3.19 is the only decision boundary that achieves maximum separation between the classes S^+ and S^- . SVMs strive to obtain this solution, which is known as the optimal hyperplane. The utilization of the optimal hyperplane enables new data points to be classified more accurately. The optimal hyperplane passes through the mid-point of classes S^+ and S^- and it ensures that the maximum distance between these classes – the maximum margin – is achieved, as illustrated in Figure 3.20.

The data points contained in S^+ and S^- which are closest to the optimal hyperplane are called support vectors. A simple rescale of \mathbf{w} for all x_i that are support vectors holds that:

$$\mathbf{w} \cdot x_i + b = 1 \quad (3.20a)$$

$$\mathbf{w} \cdot x_i + b = -1 \quad (3.20b)$$

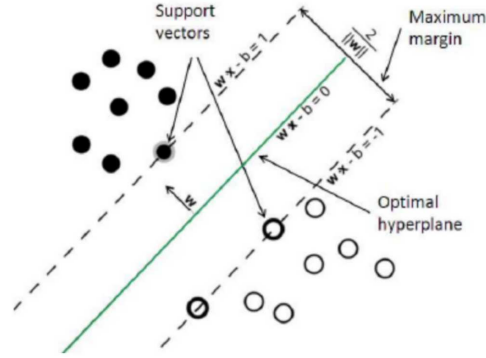


FIGURE 3.20: The optimal hyperplane separating two classes with a maximum margin [85].

The distance d between the decision boundary and the margin can be formulated as:

$$d = \frac{2}{\|\mathbf{w}\|} \quad (3.21)$$

The optimal hyperplane has the following properties: it clearly separates the data in classes S^+ and S^- ; and it achieves the maximum distance to data points in closest proximity belonging to each corresponding class. The first property states that all training data points should be classified accurately [106]. Therefore, the parameters \mathbf{w} and b of the optimal hyperplane are to be estimated, such that:

$$y_i(\mathbf{w} \cdot x_i + b) \geq 1 \text{ for } y_i = 1 \quad (3.22)$$

and

$$y_i(\mathbf{w} \cdot x_i + b) \leq -1 \text{ for } y_i = -1 \quad (3.23)$$

The combination of the two equations yields the following:

$$y_i(\mathbf{w} \cdot x_i + b) - 1 \geq 0, \forall i = 0, 1, 2, \dots, N \quad (3.24)$$

The second property states that the margin should be as large as possible. Maximizing the distance equation correspondingly minimizes $\frac{\|\mathbf{w}\|}{2}$. Thus, $f(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2$ should be minimized. Following this, the optimal hyperplane can be obtained by solving the optimization problem defined as:

$$\text{Minimize } \frac{1}{2} \|\mathbf{w}\|^2 \quad (3.25)$$

subject to

$$y_i(\mathbf{w} \cdot x_i + b) - 1 \geq 0, \forall i = 0, 1, 2, \dots, N \quad (3.26)$$

This optimization problem can be solved, given the Lagrange multipliers $\alpha_1, \alpha_2, \dots, \alpha_N \geq 0$ and the saddle point of the Lagrange function:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^N \alpha_i (y_i(\mathbf{w} \cdot x_i + b) - 1) \quad (3.27)$$

Therefore, using the Lagrange function, the optimization problem can be formulated as:

$$\text{Maximize } \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j (x_i, x_j) \quad (3.28)$$

subject to

$$\sum_{i=1}^N \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0, i = 0, 1, 2, \dots, N \quad (3.29)$$

The optimal hyperplane discriminant function under this formulation is:

$$f(x) = \sum_{i \in S} \alpha_i y_i (x_i x) + b \quad (3.30)$$

where S is the subset of support vectors corresponding to positive Lagrange multipliers.

3.3.2 Classifying Non-linear Problems

In the classification of linear problems, the classification approach simply involves the process of finding an optimal hyperplane consisting of a maximum margin which separates the data points. However, the classification of non-linear problems requires a more complex structure in order to find a hyperplane.

Non-linear problems map data points onto a higher dimensional space – the feature space – which enables the optimal hyperplane to linearly separate the data points. Figure 3.21 illustrates cases in which data points are unevenly distributed and linearly inseparable as opposed to the case in Figure 3.17.

As such, the constraint of Equation 3.24 cannot be satisfied in cases where classes are not linearly separable. To cater for such cases, a cost function is formulated which combines the margin maximization and the minimization error criteria. This solution

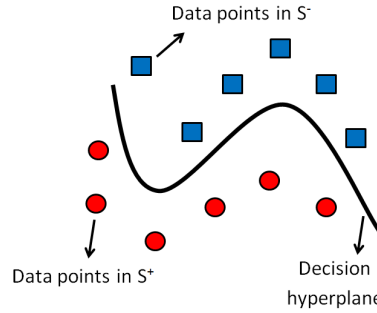


FIGURE 3.21: Data points which are not linearly separable.

involves using a set of variables, ξ_i which are known as slack variables. These variables measure the degree of misclassification. The cost function can be formulated as:

$$\text{Minimize } w, b, \xi \frac{1}{2} \| \mathbf{w} \|^2 + C \cdot \sum_{i=1}^M \xi_i \tag{3.31}$$

subject to

$$y_i(\mathbf{w} \cdot x_i + b) \geq 1 - \xi_i \tag{3.32}$$

where $\xi_i \geq 0$ and C are constants. The constant C opts for the best trade-off between the amount of error and the margin maximization. As per Mercer’s theorem [104], the dot product of the vectors in the mapping space can be equally formed as a function of the dot products corresponding to the vectors in the current space [106]. This equivalence can be expressed mathematically as:

$$\begin{aligned} K(x_i, x_j) &= \phi(x_i) \cdot \phi(x_j) \\ &= (x_i, x_i^2) \cdot (x_j, x_j^2) \\ &= x_i x_j + x_i^2 x_j^2 \\ &= x_i \cdot x_j + (x_i, x_j)^2 \end{aligned} \tag{3.33}$$

where the kernel function is represented by $K(x_i, x_j)$. This expression holds if and only if the following condition is true for any function g :

$$\int g(x)^2 dx \text{ is finite} \implies \int K(x, y)g(x)g(y) dx dy \geq 0 \tag{3.34}$$

Without prior knowledge about the explicit form of ϕ , the selection of an appropriate kernel function results in the linear separation of any data in the higher dimensional space. Therefore, the dual optimization problem can be expressed as:

$$\text{Maximize } \sum_{i=1}^M \alpha_i - \frac{1}{2} \sum_{i,j=1}^M \alpha_i \alpha_j y_i y_j (x_i, x_j) \quad (3.35)$$

subject to

$$\sum_{i=1}^M \alpha_i y_i = 0 \text{ and } \alpha \geq 0 \quad (3.36)$$

It is important to note that determining a complex curve is not suitable to separating data. Alternatively, it is possible to find an optimal hyperplane in the feature space which enables the data to be clearly separated and allow the SVM to accurately classify the unseen test data. Thus, the decision function becomes:

$$f(x) = \sum_{i \in S} \alpha_i y_i (x_i x) + b \quad (3.37)$$

where S is the set of support vectors.

3.3.3 Kernel Functions

The non-linear separability of data requires a satisfactory hyperplane for separating classes. A kernel function is employed to map the data onto higher-dimensional feature spaces. The following kernels are based on Mercer's theorem [104] and can be used by the SVM in training and classification [48]:

- Linear: $K(x_i, x_j) = (x_i)^T \cdot (x_j)$
- Polynomial: $K(x_i, x_j) = (\gamma(x_i)^T(x_j) + r)^d$, where $\gamma > 0$
- Radial Basis Function: $K(x_i, x_j) = \exp(-\gamma \cdot \|x_i - x_j\|^2)$, where $\gamma > 0$
- Sigmoid: $K(x_i, x_j) = \tanh(\gamma \cdot (x_i)^T \cdot (x_j) + r)$, where $\gamma > 0$

where r, d and γ are kernel parameters.

The prediction accuracy of the SVM corresponds to the choice of kernel, therefore, the kernel choice is crucial [22]. It should be noted that no standard method exists as a base for the selection of the most appropriate kernel [117]. However, the RBF kernel is used by many researchers and was shown to be the most accurate SVM kernel [58, 114, 117]. Therefore, the use of the RBF kernel is adopted in this research.

3.3.4 Multi-Class SVM Approaches

As explained previously, SVMs are binary classifiers that are intended towards problems involving two classes. There are currently two types of approaches used to enable SVMs to solve multi-class problems. One approach is to construct and combine several binary classifiers. The other approach is to directly consider all the data in one optimization formulation [48]. Solving multi-class SVM problems in a one step process results in variables that are proportional to the number of classes. Therefore, it is generally more computationally expensive to solve multi-class problems than to solve binary class problems.

The following subsections discuss three of the most common approaches [48].

3.3.4.1 One-against-All

This approach is known as the earliest implementation of SVMs. In this approach, M classifiers corresponding to M number of classes are constructed. Each class $i \in \{1, 2, \dots, M\}$ is separated from the data points of the remaining classes. A single class is formed by combining the data points from all classes except class i . The result is a binary classifier with a label representing class i and an additional label representing the remaining classes. This procedure is repeated M times in a rotating fashion.

In the testing phase, each of the M classifiers are presented with the input test pattern. The predicted result is determined by the i th class which obtains the maximum output value. As such, the training and testing processes are tedious due to the potentially large number of data points in each combination pair of classes.

3.3.4.2 One-against-One

This approach constructs $\frac{M(M-1)}{2}$ classifiers, each containing training data from two classes. The classifiers are combined using the Max-Wins algorithm. Each classifier is trained to distinguish between two classes using the data points in those classes as positive and negative examples.

In the testing phase, the final prediction lies in the class with the majority of votes. This voting approach is known as the Max-Wins algorithm. This results in a shorter training time than the One-against-All approach since the number of data points in the combination of classes are significantly smaller. However, its testing time, when compared to the One-against-Rest approach, is longer due to the large number of classifiers involved.

3.3.4.3 Directed Acyclic Graph

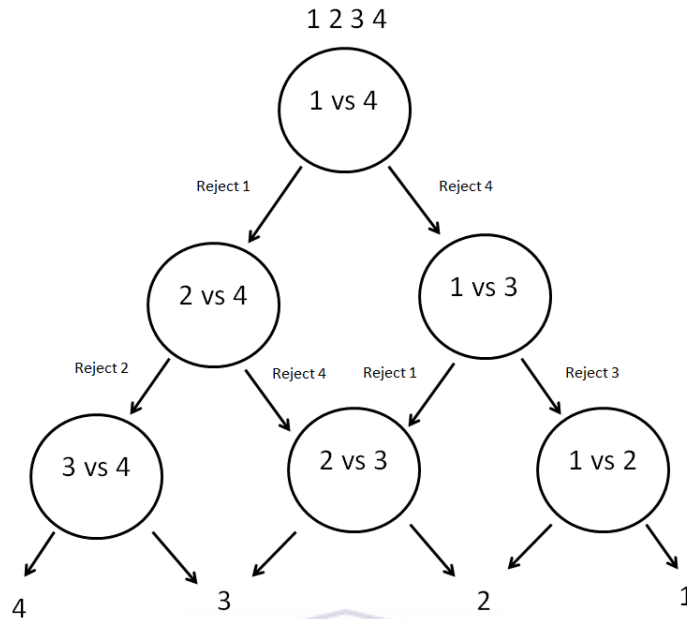


FIGURE 3.22: Directed Acyclic Graph of a 4-class problem. At each node a class is rejected until a single class remains.

The Directed Acyclic Graph (DAG) SVM algorithm was proposed by Platt *et al.* [92].

Similar to the One-against-One approach, $\frac{M(M-1)}{2}$ binary classifiers are trained using every binary pair-wise combination of the M classes. The decision strategy in the testing phase is based on a rooted binary DAG which consists of $\frac{M(M-1)}{2}$ internal nodes and M leaves, as is illustrated in Figure 3.22.

Consider a 4-class problem with $i \in \{1, 2, 3, 4\}$. Starting at the root node, classes 1 and 4 are compared. If the input pattern is classified as class 1, it simply means that class 4 was rejected. Therefore, from this node onwards, it will not be necessary to classify against class 4 again. Hence, after $M - 1 = 3$ steps, only a single predicted class will remain.

According to Platt *et al.*, an advantage of the DAG is that some analysis of generalization can be made. Furthermore, its testing time is less than the One-against-One approach.

3.4 Summary

In this chapter, the components that form part of the appearance-based FER approach proposed in this research were discussed.

The various components of the proposed face segmentation strategy were discussed. An explanation of the Viola-Jones face detection algorithm was provided and the use of the Viola-Jones algorithm for face detection in this research was justified. The skin detection technique was discussed, with much emphasis placed on the selection of an appropriate colour space as well as results for optimizing the bin width.

Two morphological operations dilation and erosion used to reduce noise and enhance features in images were discussed. A description of the eye detection strategy used was provided and justified. Subsequently, a detailed description of the LBP operator and its variants towards facial feature extraction was provided.

Finally, a detailed description behind the theory of the classification technique used by SVMs was provided.

The implementation of the techniques discussed in this chapter towards achieving FER is described in the next chapter.



Chapter 4

System Design and Implementation

This chapter focuses on the design of a facial expression recognition (FER) system using the techniques discussed in the previous chapter. Figure 4.1 illustrates the three major components of the automatic FER framework at the highest level of abstraction.



FIGURE 4.1: FER framework.

The first component involves accurately locating and isolating the face in the image. The second component involves the extraction of facial features. The final component involves the classification of facial expressions. The first three Sections 4.1, 4.2 and 4.3 of the chapter explain these three components.

Figure 4.2 illustrates a detailed high-level design of the FER algorithm. This system can automatically segment the face for both frontal images and images with any yaw rotation of the face. However, as explained in Chapter 2, the recognition of facial expressions using LBP features is angle-specific. Therefore, for the scope of this research, the feature extraction and classification components of this system are limited to frontal images and images at a yaw rotation of 60° . Since the face segmentation strategy can operate at any yaw rotation, the framework can readily be extended to FER at other rotations in future.

Section 4.4 then describes the modified, limited and constrained version of this system used to investigate the effects of various types and levels of occlusion on the FER accuracy. The chapter is then concluded.

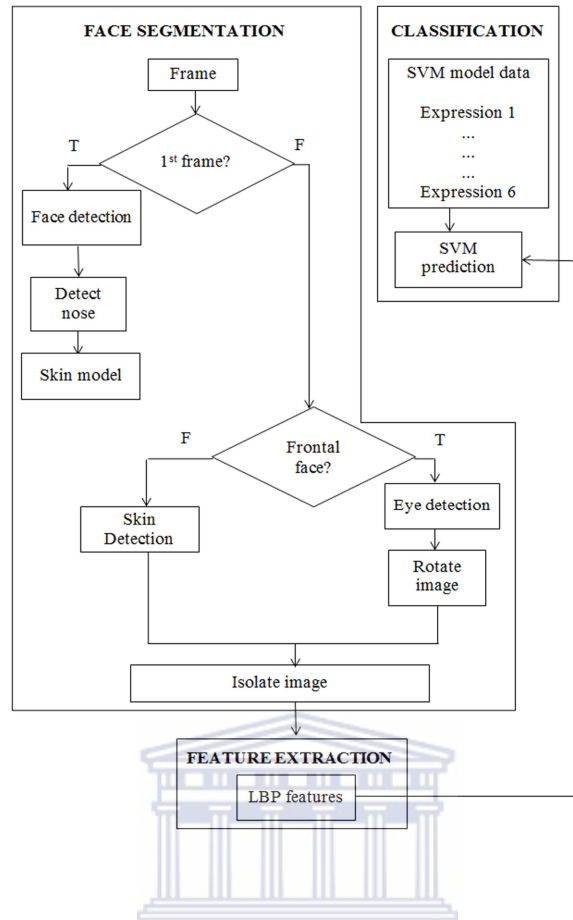


FIGURE 4.2: High-level design of the algorithm.

4.1 Face Segmentation

The face segmentation component isolates the face in an image. As explained in Chapter 1, the system works on the justified assumption that when the system is initially run, the first frame of the video sequence consists of the signer facing the camera. Using the first frame, a skin model is generated for use in isolating rotated faces.

This is done using the following procedure. The face is located using the Viola-Jones face detector. The nose region which is in the centre of the facial frame is located. Thereafter, a 10×10 pixel region around the nose is extracted, which is depicted in Figure 4.3.

A histogram of this region is computed and used as a look-up table to determine pixels that resemble the skin tone of a subject. As per Brown's bin width optimization of the skin model [13], a bin width of 8 is selected as the optimal width for skin detection.

For all other frames in the sequence, the following procedure ensues. The approach used to segment a frontal face differs from that of a rotated 1



FIGURE 4.3: Locating the nose.

whether the face is frontal or rotated depends on the result of applying the Viola-Jones face detector to the frame. If a face is detected, the frame is treated as containing a frontal face. If no face is detecting, the frame is assumed to contain a rotated face. The following sections describe the face segmentation procedure in either case.

4.1.1 Frontal Face Segmentation



FIGURE 4.4: Detecting the face.

The result of the Viola-Jones face detector is depicted in Figure 4.4. Slight tilting of the head can occur while performing facial expressions in SASL. A normalization procedure, illustrated in Figure 4.5, is applied to correct for such slight tilting. In order to clearly illustrate the normalization procedure, Figure 4.5(a) depicts an exaggerated case. In practice, the face is only expected to tilt slightly.

To overcome this, the exact positions of the eyes are obtained using the eye detection algorithm explained in the previous chapter. The result of applying this algorithm to the face in Figure 4.5(a) is illustrated in Figure 4.5(b).

Connected Component Analysis (CCA) is used to locate the coordinates (x_L, y_L) and (x_R, y_R) of the centres of the two eye blobs. These coordinates are used to calculate the angle of rotation of the head θ as follows:

$$\theta = \arctan\left(\frac{y_R - y_L}{x_R - x_L}\right) \quad (4.1)$$

This angle is used to normalize the face by aligning it with the horizontal axis by means of an affine transformation in the image. The normalized result is depicted in Figure 4.5(c).

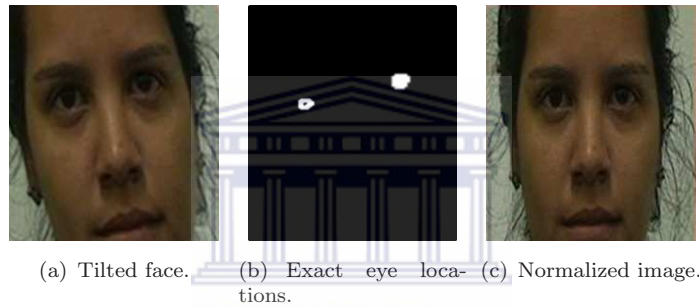


FIGURE 4.5: An example of the normalization procedure.

Note that background noise such as hair and ears may still be visible in this image. In order to more accurately isolate the face, the Viola-Jones algorithm is used with an eye detection cascade to detect the eye region. The result of applying this algorithm is illustrated in Figure 4.6.



FIGURE 4.6: Detecting the eye region.

The width of the detected eye region in Figure 4.6 is used together with the height of the facial region in Figure 4.4 to isolate the face and completely remove background noise, as illustrated in Figure 4.7.



FIGURE 4.7: Isolated frontal face.

4.1.2 Rotated Face Segmentation

For rotated faces, the skin model histogram computed earlier on is backprojected onto the image producing a greyscale image in which the skin regions are emphasized. A threshold value of 60, as per Achmed [3], Li [65] and Brown's [13] work, is used to create a binary image in which the skin is represented by white pixels and non-skin pixels are represented by black pixels. A highly rotated face and the result of applying skin detection to the face are illustrated in Figure 4.8.

The two morphological operations erosion and dilation, in that order, are applied to the resulting image to remove excess noise and restore discontinuities in the skin region. The result of applying these operations is illustrated in Figure 4.9. The most important point to note from the figure is that the largest skin blob contains no discontinuities or holes.

CCA is applied to the resulting skin image in order to locate all of the skin blobs in the image. The largest skin blob is considered to be the face. All other contours are eliminated from the image. The resulting contour is illustrated in Figure 4.10. It is clear that this contour accurately isolates the face region.

The coordinates of the top-most, bottom-most, left-most and right-most extents of this contour map out a rectangle which contains the isolated rotated face in the original image, as illustrated in Figure 4.11.

4.2 Feature Extraction

The Local Binary Pattern (LBP) operator is applied to the isolated frontal or rotated facial images, as illustrated in Figure 4.12. It was shown in Chapter 2 that the resolution of the facial image before applying the LBP operator directly affects the FER accuracy. The optimal resolution was also shown to be angle-specific. Facial images at different rotations achieved optimal performance at varied resolutions. Therefore, an



(a) Original rotated facial image.



(b) Skin detection result for the rotated facial image.

FIGURE 4.8: Skin detection result for rotated images.

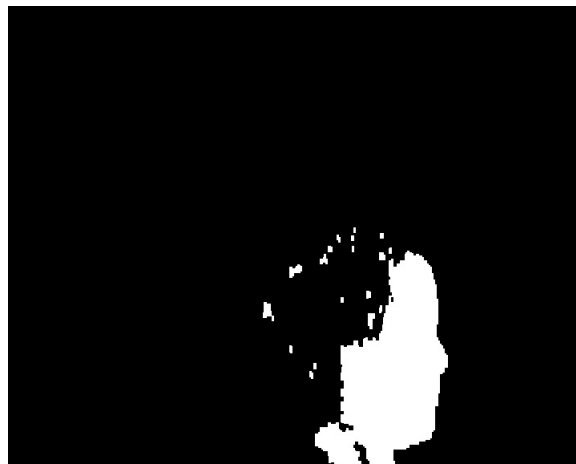


FIGURE 4.9: Skin image with morphological operations.

experiment was carried out to determine the optimal resolution for the frontal and 60° case. Additionally, the number of regions into which the LBP image is divided and used



FIGURE 4.10: Detecting the contours of the face.



FIGURE 4.11: Isolated rotated face.

to compute histograms – the region size – is expected to affect the FER accuracy. Therefore, in addition to optimizing the resolution size, the region size was also optimized, which has not been carried out in the literature to our knowledge.

As illustrated in the related work chapter, the process of determining an optimum resolution size is one of trial and error. Combinations of varied image height and width are considered and the cross-validation accuracy of the SVM trained on $(n-1)$ subsets of a training set and tested on 1 subset is used as a measure of optimality of the resolution. As explained in the previous chapter, the values of C and γ of the SVM can affect the resulting accuracy. Therefore, the experiment used to optimize the resolution and region size went hand-in-hand with the optimization of the C and γ parameters. This experiment is described in the next chapter.

At this stage, it suffices to say that, for frontal images, the optimal resolution size was 40×60 with a region size of 8×10 . For rotated images, the optimal resolution size was 40×50 with a region size of 8×5 .

The facial image resulting from the procedure in the previous sections in each case is scaled to the optimal resolution.

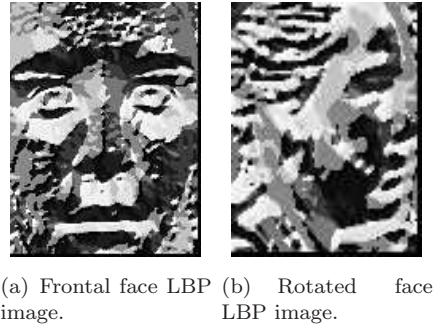


FIGURE 4.12: Applying the LBP operator.

The LBP operator $LBP_{8,2}^{u2}$ used is a combination of the uniform LBP operator LBP^{u2} and the extended LBP operator with a pixel neighbourhood of 8 pixels and a radius of 2 pixels $LBP_{8,2}$. These two operators were shown, amongst the purely LBP operators, to be the most accurate in a previous Chapter 2. A combination of these operators is expected to provide highly accurate results with the advantage of the efficiency provided by the uniform operator.

The $LBP_{8,2}^{u2}$ operator is applied to the scaled isolated facial image. The resulting image is divided into the optimal number of regions, as illustrated in Figure 4.13.

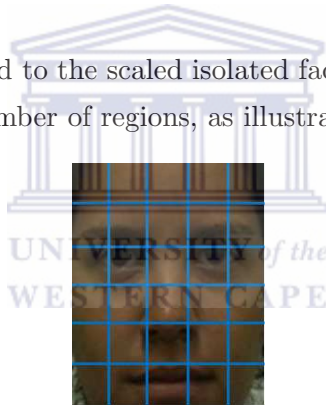


FIGURE 4.13: Facial image divided into regions.

A histogram of each region is computed which encodes a frequency count of each binary pattern. As explained in the previous chapter, the histogram contains 59 bins. Finally, the region histograms are concatenated into a single spatially enhanced histogram. The spatially enhanced histogram is used as a feature vector in the training and testing phases which are explained in the following section.

Assuming the optimal number of region histograms per image to be h , optimized in the next chapter, and 59 bins per histogram yields a feature vector length of $h \times 59$. A predefined label $L \in \{1, 2, \dots, 6\}$ corresponding to each facial expression class is assigned to each feature vector. With this specification, the feature vector can be expressed mathematically as:

$$V = \{(L, H_r)\} \quad (4.2)$$

where H_r is the histogram at region $r \in \{0, 1, \dots, (h - 1)\}$. An example of a possible training data file is illustrated in Figure 4.14. The index in the top-left corner of the file is the class label. Subsequent items are features in the format $i : v$ where i is the feature index and v is the value in floating-point notation. For example, in the file illustrated, the 1 in the top-left corner represents a facial expression class of 1 and the item right of the class label (1 : 3.000000) represents feature index 1 with a value of 3.000000. The data is scaled to avoid features with high numeric values from dominating features with low numeric values [48].

```

1 1:3.000000 2:0.000000 3:4.000000 4:9.000000 5:0.000000 6:0.000000 7:1.0000
11:0.000000 12:1.000000 13:0.000000 14:0.000000 15:0.000000 16:0.000000 17:0
20:0.000000 21:0.000000 22:0.000000 23:0.000000 24:0.000000 25:0.000000 26:0
29:0.000000 30:0.000000 31:0.000000 32:1.000000 33:0.000000 34:66.000000 35:
38:0.000000 39:0.000000 40:0.000000 41:139.000000 42:45.000000 43:7.000000 4
47:0.000000 48:0.000000 49:1.000000 50:0.000000 51:0.000000 52:0.000000 53:1
56:0.000000 57:0.000000 58:0.000000 59:0.000000 60:2.000000 61:3.000000 62:1
65:0.000000 66:0.000000 67:7.000000 68:13.000000 69:0.000000 70:0.000000 71:
74:0.000000 75:1.000000 76:5.000000 77:2.000000 78:0.000000 79:0.000000 80:0
83:2.000000 84:0.000000 85:0.000000 86:0.000000 87:0.000000 88:0.000000 89:0
92:0.000000 93:4.000000 94:47.000000 95:48.000000 96:23.000000 97:6.000000 9
101:73.000000 102:35.000000 103:5.000000 104:1.000000 105:5.000000 106:21.00
109:3.000000 110:0.000000 111:1.000000 112:2.000000 113:0.000000 114:0.00000
117:1.000000 118:0.000000 119:1.000000 120:3.000000 121:6.000000 122:8.00000
125:0.000000 126:0.000000 127:2.000000 128:4.000000 129:0.000000 130:0.00000
133:0.000000 134:0.000000 135:0.000000 136:86.000000 137:20.000000 138:0.000
141:29.000000 142:20.000000 143:0.000000 144:1.000000 145:0.000000 146:0.000

```

FIGURE 4.14: Example SVM training file.

4.3 Training and Testing Phases

This section discusses the procedure involved in training on, and classification of, the prototypic facial expressions using an SVM. The SVM is trained on a set of data described in Section 4.3.1 in the training phase which is discussed in Section 4.3.2. The training procedure involves determining the optimum C and γ parameters of the SVM for the data set. However, the optimization of these parameters for the proposed feature extraction process goes hand-in-hand with determining an optimum resolution and region size for the feature vector. Therefore, the experiment used to determine the optimum C and γ parameters along with the feature vector optimization procedure is described in detail in the next chapter.

Once the SVM is trained, any unseen image can be used as input to the system for classification in the testing phase. The process involved in classifying an unseen image in the testing phase is discussed in Section 4.3.3.

4.3.1 Training Set

The Binghamton University 3D Facial Expression (BU-3DFE) database was used for training and testing. The database contains 100 subjects (56 female, 44 male), ranging

from 18 years to 70 years old, with various ethnicities and skin tones.

In the collection of the data set, each subject performed seven expressions – “Anger”, “Disgust”, “Fear”, “Happiness”, “Sadness”, “Surprise” and “Neutral” – in front of a 3D scanner. For each of the six prototypic expressions (which excludes the “Neutral” expression), four levels of intensity for each expression were captured ranging from 1 – 4 where 4 was the peak of the expression and 1 was close to but not “Neutral”. 3D models of each subject performing each expression and intensity level were captured. The raw data takes the form of a face texture and a 3D model. In order to obtain images of each (subject, expression, intensity) combination, a 3D graphics tool can be used to perform UV mapping of the texture to the model to obtain an accurate depiction of the original subject’s head. This can be rotated to any desired angle. The resulting image can be rendered as a 2D image. In this research, the open source graphics tool Blender [9] was used for this purpose. This was used to obtain frontal images and images rotated to 60°.

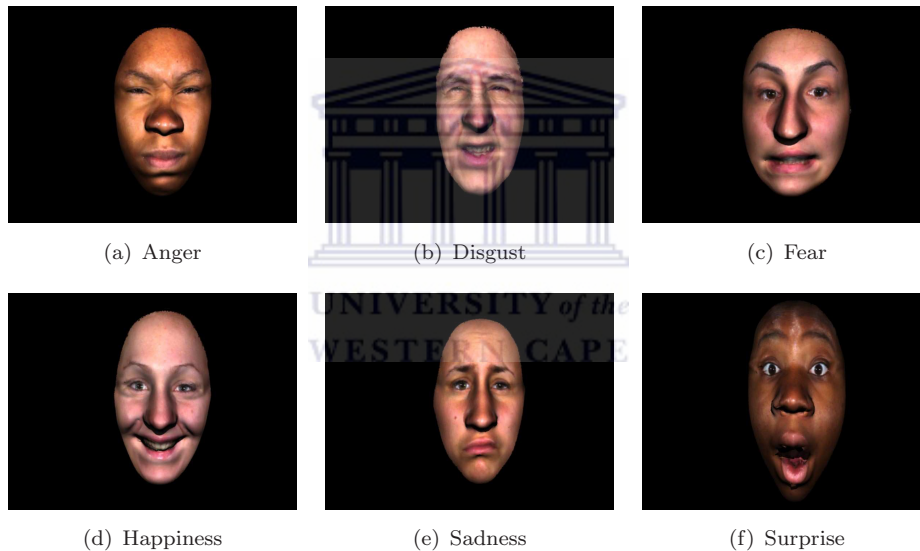


FIGURE 4.15: The six prototypic facial expressions used from the BU-3DFE database [118].

This research only focuses on the recognition of the six prototypic expressions and on the images of the highest intensity level – the peak of the expression. Figure 4.15 depicts an example of the six prototypic expressions for the frontal and rotated case from the BU-3DFE database. The images of a total of 10 subjects from the data set were used for training. This resulted in 10 frontal and 10 images rotated to 60° per expression, an overall total of 60 frontal and 60 rotated images for training. This set is henceforth referred to as the “training set”.

4.3.2 Training Phase

As explained in Chapter 2, the recognition accuracy of LBPs is angle-specific. Therefore, Moore and Bowden [77] developed a pose classifier which identifies at which angle the face is rotated prior to selecting the appropriate classifier for training. Therefore, a similar approach is adopted in this research. Two separate classifiers, one for the frontal and one for the rotated case, are used.

The training procedure is illustrated in Figure 4.16. This procedure is carried out separately for each classifier. For the frontal classifier, the feature vector corresponding to each frontal image in the training data set is computed. The feature vector is labelled with the corresponding prototypic expression label. The entire training set is scaled in order to avoid features with large values from dominating features with smaller values in the feature set.

The LibSVM grid search function uses cross-validation to obtain optimum C and γ values which can be used as parameters to train an SVM. Cross-validation partitions the data set into v equal subsets, where $v - 1$ subsets are used for training and the remaining subset is used for testing. This procedure is repeated v times in a rotating fashion, such that different training and testing sets are selected in each case. For each training and testing set combination, the procedure selects a different combination of C and γ values and computes a cross-validation accuracy. The C and γ values which correspond to the highest cross-validation accuracy are deemed optimal and are used to train the SVM. When using optimum C and γ values to train an SVM, the trained SVM is thus optimized. This procedure is repeated for the rotated classifier.

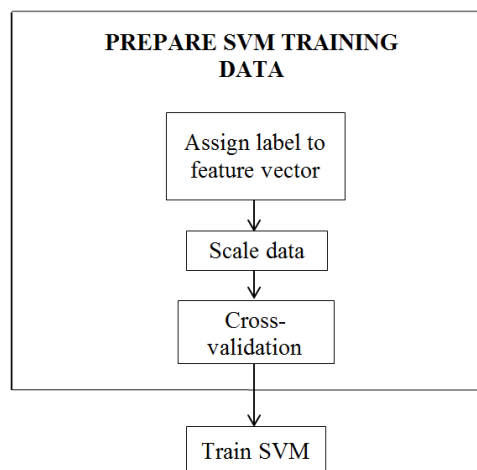


FIGURE 4.16: SVM training procedure.

The next chapter details the experimentation carried out to optimize the SVMs. At this stage, it suffices to state that, for frontal images, the optimum resolution size was

40×60 at an optimum region size of 8×10 . With these sizes, an optimum accuracy of 72.67% was obtained with $C = 0.5$ and $\gamma = 0.0078125$. For images rotated to 60° , the optimum resolution size was 40×50 at an optimum region size of 8×5 . With these sizes, an optimum accuracy of 69.67% was obtained with $C = 8.0$ and $\gamma = 3.0517578125e^{-05}$. With these parameter values, two new separate SVMs were trained, one for the frontal and one for the rotated images on the corresponding images of the training set.

4.3.3 Testing Phase

The testing phase involves carrying out classification with the trained model on a previously unseen image. This procedure is illustrated in Figure 4.17.

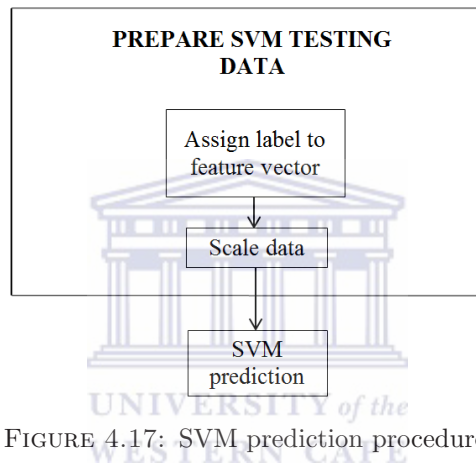


FIGURE 4.17: SVM prediction procedure.

The feature extraction procedure is repeated for the unseen image and the feature vector data file is created in the same manner. However, a default label of 0 is assigned since the correct label is not known. This feature vector is used as input to the SVM which predicts the class to which the given feature vector belongs. As before, this class corresponds to a particular facial expression in the set $L \in \{1, 2, \dots, 6\}$.

4.4 Simulating Occlusion

As set out in Chapter 1, the proposed fully automatic FER system illustrated in Figure 4.2 is not expected to discern between whether or not the face is occluded. The fully automatic FER system uses the Viola-Jones face detector to distinguish between a frontal and rotated face. However, if the Viola-Jones face detector fails to detect a frontal face, as is the case with an occluded image, the image is then treated as containing a rotated face. As such, it is not then possible to obtain results for frontal occluded images.

Therefore, in order to investigate the effects of occlusions on frontal and rotated images, two limited versions of the fully automatic system were created. One system deals with the case of frontal occluded images only and expects such images as input. The other deals with the case of rotated occluded images and only expects such images as input. The system dealing with frontal facial occlusion uses the SVM model for frontal images. The system dealing with rotated facial occlusion uses the SVM model for rotated images.

Figures 4.18 and 4.19 depict the systems which take in occluded frontal and rotated images as input, respectively.

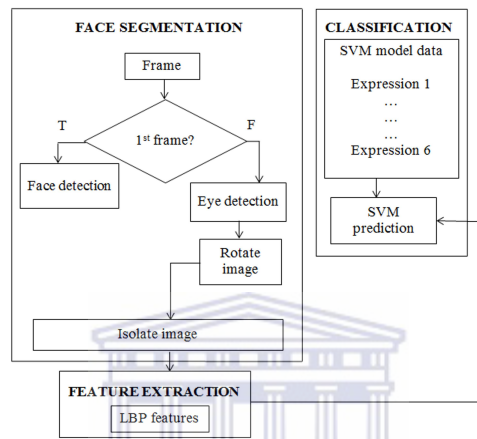


FIGURE 4.18: System for frontal occluded images.

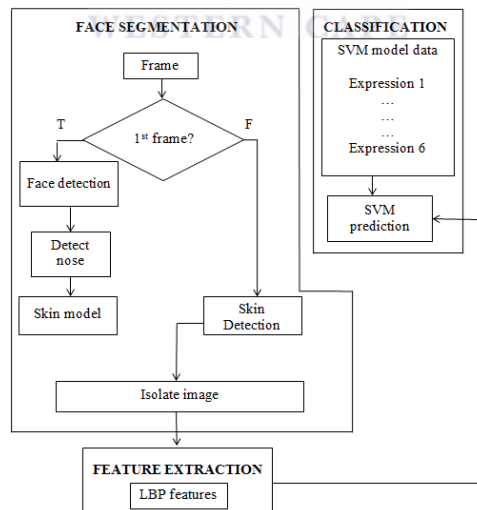


FIGURE 4.19: System for rotated occluded images.

4.4.1 Summary

In this chapter, the implementation of the fully automatic FER system was discussed. The system contains three major components and the implementation of each individual

component was explained. The face segmentation procedure was illustrated and shown to be a highly successful procedure for both frontal images and rotated images. The LBP operator was implemented and the resulting feature vectors were discussed. The optimization of the resolution size of the image prior to LBP computation and the region size for histogram computation was carried out. It was shown that a resolution of 40×60 and region size of 8×10 was optimum for frontal images. For rotated images, the optimum resolution and region size differed. In this case, a resolution of 40×50 at an optimum region size of 8×5 was optimum. Two SVMs were trained, one for the frontal and one for the rotated case. Finally, two limited versions of the fully automatic FER system were created for frontal and rotated occluded images in order to investigate the effects of partial occlusions on FER using LBPs.

The next chapter discusses the experiments carried out using this system to assess the FER and face segmentation accuracy.



Chapter 5

Experimental Results and Analysis

This chapter presents the assessment of the robust FER (FER) system. The two databases used in experimentation are described. The Binghamton University 3D Facial Expression (BU-3DFE) database is used for testing the face segmentation accuracy as well as the recognition accuracy of facial expressions.

In addition to the BU-3DFE Database, a database containing five subjects in a complex background is generated to test the face segmentation procedure in a complex background since the BU-3DFE database only contains facial images in a simple background.

For FER experiments, the criterion is explained and the outputs of the system are evaluated. Experimental analysis is performed on the effectiveness of the recognition procedure.

The aim of the face segmentation experiment is to illustrate that the system is able to accurately segment a face in a complex background. The FER experiments are aimed at determining the success rate of the system by evaluating how rotation and partial occlusions of the face affect the system.

All experiments were carried out on a PC containing an Intel i7 3770k 3.5 GHz quad core CPU, an NVIDIA 580GTX GPU and 16 GB RAM, running the Kubuntu 11.04 x64 operating system.

The rest of the chapter is organized as follows: Section 5.1 discusses accuracy testing for face segmentation; Section 5.2 discusses the feature vector and optimization of the Support Vector Machine; and Section 5.3 discusses accuracy testing for FER.

For the face segmentation experiment, a different dataset containing five subjects of various ethnicities and skin tones were used. A Logitech C910 web camera was used at a resolution of 640×480 pixels at a frame rate of 15 frames per second (FPS).

5.1 Face Segmentation Experiment

This section describes the experiment carried out in order to assess the accuracy of the face segmentation procedure under varied conditions including varied subject skin tones and on a complex background. This experiment aims to answer the first research question posed in Chapter 1: “Can the proposed face segmentation strategy accurately segment the face in facial images with varied skin tone, in the presence of rotations and on a complex background?”. Therefore, the analysis focuses on the effects of three factors on the segmentation accuracy: subject skin tone, rotation and complex background.

The subsections that follow describe the data set used in this experiment, the criterion used to judge the accuracy of a segmented face, the exact experimental procedure, the results that were obtained and an analysis of the results to answer the research question.

5.1.1 Data Set

For this experiment, 5 South Africans with diverse skin tones were used. Each subject was required to sit on a chair facing the web camera and instructed to continuously rotate their heads from side to side, up to and including an angle of 90° on either side, for a total of five seconds. This was illustrated to each subject prior to video capturing. Using a web camera at a frame rate of 25 FPS, this resulted in a database containing a total of 625 images at a variety of rotations of the head across all subjects. This data set is henceforth referred to as the “tracking data set”. Figure 5.1 illustrates the five subjects. It is clear that the subjects contain different skin tone and that each subject is on a slightly different background.

In addition, two images of 50 subjects, one frontal and one rotated to 60° , from the BU-3DFE data set explained in the previous chapter, were also used for this experiment. The data set consists of a total of 300 frontal images and 300 images rotated to 60° across 50 subjects of varied skin tone on a simple background.

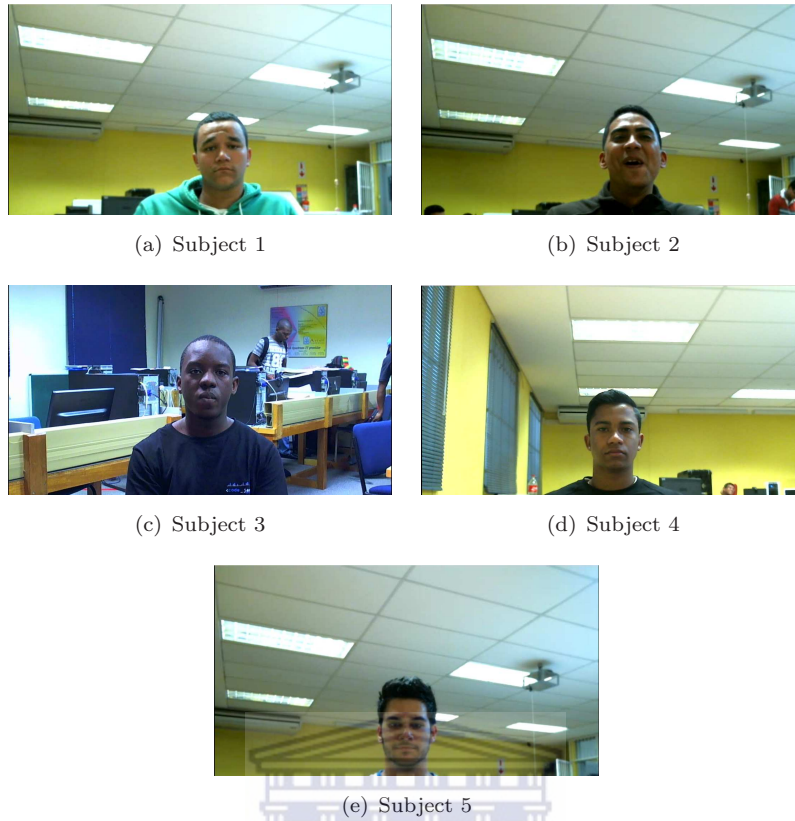


FIGURE 5.1: An example of the five subjects each on a slightly different complex backgrounds.

5.1.2 Experimental Procedure

All the images from the tracking data set and images of 50 subjects from the BU-3DFE data set were used as input to the face segmentation procedure of the proposed system discussed in Chapter 4. In each case the system produced a segmented output corresponding to the input frame. The criterion for an accurately segmented face, explained in the next subsection, was used to determine the outcome. The results were recorded and analyzed.

5.1.3 Criterion for an Accurately Segmented Face

Each input frame and the resulting segmented output frame, which is expected to consist of the segmented face only, are compared. Similar to the procedures used by Kolsch and Turk [60] and Li [65], a visual comparison between the input frame and the resulting frame is carried out. An accurately segmented face is considered as a face which does not contain any extra information surrounding the face such as hair on the facial sides etc. but does not crop any features of the face such as the facial sides, mouth or forehead out either. With regards to the forehead, research shows that only the region

immediately above the eyes is necessary for FER [29, 105]. Therefore, having a portion of the forehead above the eyes is sufficient, but having the entire forehead is also considered as an accurately segmented face. An incorrectly segmented face has the following characteristics:

- Hair on the facial sides beyond either ear present.
- Part or all of the neck present.
- Hair on top of the forehead present.
- Part or all of either eye, the mouth or the facial sides cut off.
- The entire forehead cut off.

Therefore, the criterion for an accurately segmented face is one that is not an incorrectly segmented face as per the above definition. Figure 5.2 illustrates examples of accurately segmented frontal and rotated faces for one subject. Similar to Kolsch and Turk [60] and Li [65], the researcher carried out these comparisons.



(a) Face rotated to the right. (b) Frontal face.



(c) Face rotated to the left.

FIGURE 5.2: Accurately segmented frontal and rotated faces for one subject.

5.1.4 Results and Analysis

Table 5.1 indicates the number of frames in which the face was correctly segmented compared to the total number of frames for the tracking data set.

The face segmentation procedure achieved an average recognition accuracy of 97.1%. The results indicate a near-perfect segmentation accuracy. It is clear that the procedure

TABLE 5.1: Face segmentation accuracy for the tracking data set.

Subject	Correct	Total	Accuracy (%)
1	125	125	100.0
2	125	125	100.0
3	107	125	85.6
4	125	125	100.0
5	125	125	100.0

is highly consistent achieving an accuracy of 100% for all subjects, with the exception of Subject 3 who registered an accuracy of 85.6%. Noting that the subjects were of completely varied skin tones, on complex backgrounds and rotating their heads, these results are indicative of a robust face segmentation procedure.

The results in Table 5.1 indicate that for 18 of the 125 cases for Subject 3, the face segmentation procedure detected an object situated to the right of the subject's face, as illustrated in Figure 5.1(c).

Since this procedure depends primarily on skin colour distribution to accurately carry out face segmentation, a colour distribution analysis was carried out. Figure 5.3 illustrates the same histogram computed for both the skin distribution of the subject and the colour distribution of the detected object incorrectly perceived as the face. A similarity between the histogram of the skin model – the face – and the histogram of the detected object, makes it difficult for the procedure to accurately distinguish between the face and an object in close proximity to the face.

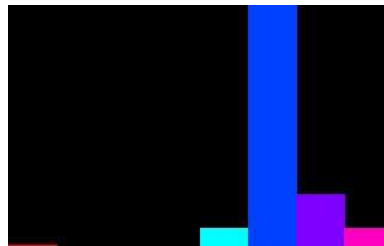


FIGURE 5.3: The same histogram computed for the face and the detected object incorrectly perceived as the face.

Based on the observation of the results obtained for Subject 3, an improvement was made to the face segmentation procedure to cater for this possible source of error. The subject is assumed to be stationary, with only natural movements of the body and the head rotating from side to side. As explained in Chapter 1, the first frame of the video sequence is assumed to contain a frontal face. Based on these two assumptions, after detecting the face in the first frame, in all subsequent frames, a boundary region with a size of 20% larger than the face is used as a location threshold for detecting the face.

The procedure does not attempt to detect a face outside this boundary. This can limit potential sources of noise.

The above experiment was repeated with the improved face segmentation strategy and achieved a perfect 100% face segmentation accuracy for all subjects. Thus, a face segmentation strategy that is invariant to the skin tone of subjects and a complex background was developed.

For the BU-3DFE data set, the system consistently registered a perfect face segmentation accuracy of 100% for all 50 subjects for, both, the frontal and the rotated positions. It is important to note that the 50 subjects in this data set are of varied skin tone. This clearly demonstrates, once again, that the strategy is invariant to the skin tone of subjects.

Therefore, the response to the research question posed is: The proposed face segmentation strategy can accurately segment the face in facial images with varied skin tone, in the presence of rotations and on a complex background. The strategy is invariant to the skin tone and a complex background.

5.2 Feature Vector and SVM Optimization

This section discusses the procedure that was followed in order to optimize the resolution and region size of frontal and rotated images. This process went hand-in-hand with the optimization of the SVM using LibSVM's grid search function. The optimum values obtained from the grid search function are given as well as the optimum resolution and region size for frontal and rotated images.

The subsections that follow describe the exact experimental procedure, the results that were obtained and an analysis of the results.

5.2.1 Experimental Procedure

To optimize the resolution R and region size G , combinations of varied resolution and region size (R, G) were used to generate a feature vector, as previously explained. The grid search function in LibSVM [17] was used to determine the C and γ values that yield the highest cross-validation accuracy for the (R, G) combination. The search investigates the accuracy of a number of possible C and γ values exhaustively and selects the pair that achieves the highest cross-validation accuracy. Cross-validation divides the training set into v equally-sized subsets, where the classifier is trained on $v - 1$ subsets and tested on the remaining subset [48]. The cross-validation accuracy is the average accuracy across all v combinations.

As explained in the related work chapter, the process of determining an optimal resolution is a process of trial and error. Numerous possibilities exist. Attempting to optimize the region size in addition to this further adds to the number of possibilities. For the scope of this research, width and height combinations of 40, 50 and 60 pixels for the resolution size and 5, 8 and 10 pixels for the region size were considered. In cases where the resolution dimension was not a multiple of the region dimension, such as a resolution width of 50 and a region width of 8, the combination was ignored. This procedure was carried out separately for the frontal images and images rotated to 60° .

5.2.2 Results and Analysis

Table 5.2 summarize the results obtained for the frontal images and 5.3 summarize the results for the images rotated to 60° .

TABLE 5.2: Optimized resolution and region size for frontal images.

R=(40 × 40)				R=(40 × 50)				R=(40 × 60)			
G	5	8	10	G	5	8	10	G	5	8	10
5	69.33	68.33	67.33	5	71	-	69	5	69	-	70
8	69.67	71.33	68.33	8	72.33	71.33	-	8	71.67	-	72.67
10	71.67	71	71	10	71	-	72	10	72.33	-	71.67

R=(50 × 40)				R=(50 × 50)				R=(50 × 60)			
G	5	8	10	G	5	8	10	G	5	8	10
5	56.67	64.67	62.67	5	53	-	50.67	5	52	-	45.67
8	-	-	-	8	-	-	-	8	-	-	-
10	64.33	65	63	10	50.67	-	61	10	47.67	-	60

R=(60 × 40)				R=(60 × 50)				R=(60 × 60)			
G	5	8	10	G	5	8	10	G	5	8	10
5	58	54.67	53.67	5	50.67	-	50	5	47	-	42.33
8	-	-	-	8	-	-	-	8	-	-	-
10	50.67	64.67	62.33	10	44	-	37	10	41	-	37

For frontal images, the optimum resolution size was 40×60 at an optimum region size of 8×10 . With these sizes, an optimum accuracy of 72.67% was obtained with $C = 0.5$ and $\gamma = 0.0078125$. For images rotated to 60° , the optimum resolution size was 40×50 at an optimum region size of 8×5 . With these sizes, an optimum accuracy of 69.67% was obtained with $C = 8.0$ and $\gamma = 3.0517578125 \times 10^{-05}$.

With these parameter values, two new separate SVMs were trained, one for the frontal and one for the rotated images. The frontal SVM was trained with its corresponding

TABLE 5.3: Optimized resolution and region size for rotated images.

R=(40 × 40)				R=(40 × 50)				R=(40 × 60)			
G	5	8	10	G	5	8	10	G	5	8	10
5	64.33	65.33	64	5	66.67	-	66	5	68	-	63.33
8	66.67	66	65.33	8	69.67	-	67.67	8	67.67	-	65
10	65.33	65	65.33	10	68.67	-	64.33	10	67	-	64.67

R=(50 × 40)				R=(50 × 50)				R=(50 × 60)			
G	5	8	10	G	5	8	10	G	5	8	10
5	59.33	53.67	58	5	58.67	-	52.67	5	61.33	-	54
8	-	-	-	8	-	-	-	8	-	-	-
10	56.67	54.33	58	10	59	-	51.67	10	58.67	-	51

R=(60 × 40)				R=(60 × 50)				R=(60 × 60)			
G	5	8	10	G	5	8	10	G	5	8	10
5	55	52.3	51.67	5	57.33	-	48.67	5	58.33	-	48
8	-	-	-	8	-	-	-	8	-	-	-
10	54.67	49	50.33	10	54.67	-	46.67	10	53.33	-	47.33

parameters using images of 10 of the subjects in the frontal pose from the BU-3DFE data set – a total of 10 images, as explained in the previous section. The rotated SVM was similarly trained with its corresponding parameters using images of 10 of the subjects in the rotated pose from the BU-3DFE data set – a total of 10 images.

5.3 Facial Expression Recognition Accuracy Testing

This section describes the experiment carried out in order to answer the second research question posed in Chapter 1: “Can whole facial expressions be recognized at a high accuracy using the LBP operator in the presence of rotations and partial occlusions of the face?”. The analysis of this question can be broken down into an investigation of the FER accuracy of the following four categories of faces:

1. Frontal faces – frontal faces without any occlusions.
2. Rotated faces – faces rotated to 60° without any occlusions.
3. Frontal occluded faces – Frontal faces with occlusions.
4. Rotated occluded faces – Rotated faces with occlusions.

The experiments for all these categories made use of a single criterion to determine a correctly recognized facial expression. This criterion is described in Section 5.3.1. Thereafter, the experiments carried out to assess the FER accuracy for each of the four categories of faces, with a subsequent analysis, are described in the sections that follow.

5.3.1 Criterion for a Correctly Recognized Facial Expression

The system aims to accurately recognize each of the six prototypic facial expressions. Each image in the database was labelled as one of the six prototypic facial expressions. In each case, the system response for each input frame was compared to the ground truth. The system classifies each input frame as one of the six prototypic facial expressions. If the output of the system for a particular input frame matches its corresponding label in the database, it is deemed a correct classification. Otherwise, it is deemed an incorrect classification.

5.3.2 Frontal Facial Images

This section discusses the experimental procedure carried out for frontal facial images and provides an analysis of the results obtained for each experiment.

5.3.2.1 Experimental Procedure

Frontal facial images of 40 subjects from the BU-3DFE data set, not in the training set, were used as input to the system. These images contain subjects facing the camera. In each case, the output of the system was analyzed using the criterion for accurately recognizing a facial expression.

5.3.2.2 Results and Analysis

For reference, the full set of results obtained for the 40 subjects is provided in Table A.1 in Appendix A. The table provides the system response for each image of each subject used as input. Table 5.4 is a confusion matrix summarizing the results after applying the FER strategy to frontal facial images. Table 5.5 summarizes the results as a percentage per expression.

The system registered an average accuracy of 75% for frontal facial images ranging from 62% to 90%. It is noted that “Surprise” registered the highest accuracy of 90%, and the lowest, but by no means low, accuracy of 62% was registered by “Disgust” and

TABLE 5.4: Confusion matrix for frontal FER accuracy.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	33	4	1	0	2	0
Disgust	8	25	5	1	0	1
Fear	1	3	25	7	1	3
Happiness	1	0	4	35	0	0
Sadness	6	0	5	0	28	1
Surprise	1	1	2	0	0	36

TABLE 5.5: Average Frontal FER accuracy.

Expression	Correct (40)	Average (%)
Anger	33	82
Disgust	25	62
Fear	25	62
Happiness	35	87
Sadness	28	70
Surprise	36	90

“Fear”. It should be noted that the system achieved a high accuracy of above 80% for three of the expressions, above 70% for four of the expressions and above 60% over all expressions. These results are very encouraging and are indicative of a highly successful feature extraction process. In terms of responding to the research question, it is clear that the proposed system can recognize whole expressions at a high accuracy in frontal non-occluded images.

Comparing these results with Moore and Bowden’s [77] results indicates that the system achieves a higher average recognition accuracy of 75% using the combined $LBP_{8,2}^{u2}$ operator than Moore and Bowden’s accuracies of 72% and 62% for LBP^{ms} and LBP^{u2} , respectively. However, this result is only indicative since Moore and Bowden trained and tested on the entire data set – 100 subjects. In future, the testing can be extended to the entire data set to perform a direct comparison. At this stage, it is only possible to say that the result indicates that the combined operator may be a better facial expression descriptor than the individual operators.

Analyzing the results to identify possible sources of error in the two lowest performing expressions in Table 5.4 indicates that “Disgust” was misclassified as “Anger” in the majority of incorrect classifications – 8 cases – and “Fear” was misclassified as “Happiness” in the majority of incorrect classifications – 7 cases. The expression “Disgust” is performed by frowning, tightly pursing the lips and flaring the nostrils and “Anger” is performed by frowning and tightly pursing the lips only, it was initially expected that in

some cases these two expressions may be confused with each other due to resemblance. Analyzing the table, it is in fact seen that in the majority of errors for the expression “Anger”, “Anger” was similarly confused with “Disgust”. Figure 5.4 depicts an example of an expression “Disgust” that looks similar to “Anger”.

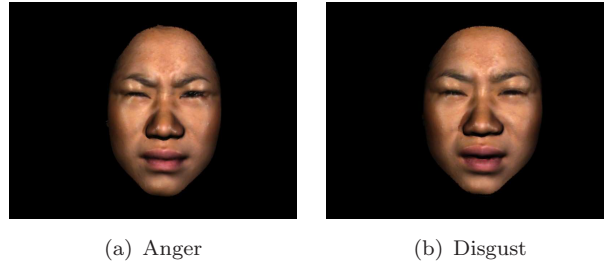


FIGURE 5.4: An example of a case in which “Anger” was expressed similarly to “Disgust” [118].

Similarly, “Fear” and “Happiness” are both primarily performed with the cheek regions by stretching out the mouth. Therefore, it was initially expected that in some cases these two expressions may be confused with each other, depending on how they are performed by individual subjects. In fact, the results indicate that in most cases “Happiness” was similarly confused with “Fear”. Figure 5.5 depicts an example of an expression “Fear” that looks similar to “Happiness”.

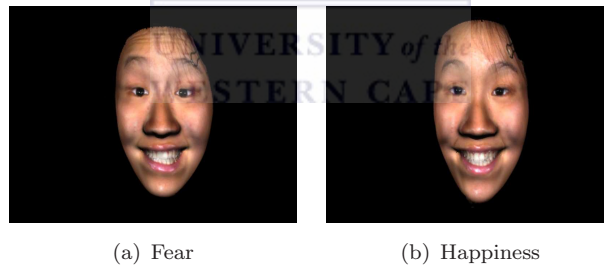


FIGURE 5.5: An example of a case in which “Fear” was expressed similarly to “Happiness” [118].

Figure 5.6 graphically summarizes the FER accuracy per subject. For convenience, the results in the figure are sorted in descending order. For reference, the full set of results is provided in Table A.2 in Appendix A. The results indicate that for 33 out of the 40 subjects – 82% of the subjects – the FER system correctly recognized at least 4 out of the 6 expressions. Furthermore, for 38 out of the 40 subjects, the FER system correctly recognized at least 3 out of the 6 expressions. These results are very encouraging and indicative of a robust FER strategy which is subject invariant and which generalizes very well to a large group of subjects with different skin tones, gender and face dimensions. For only two subjects, the system recognized less than 3 expressions but it is important to note that, for these subjects, the recognition rate was at least 1 out of 6.

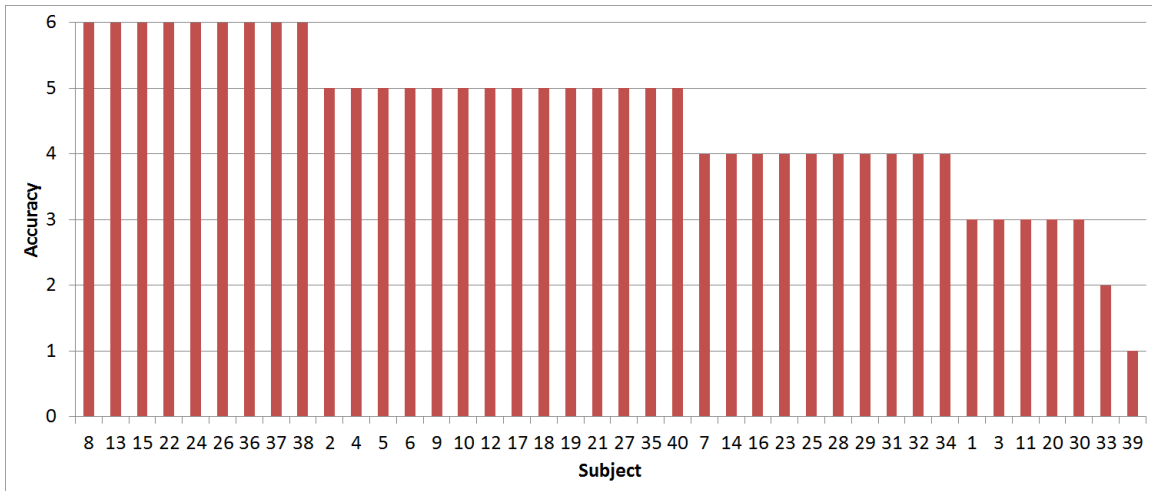


FIGURE 5.6: FER accuracy per subject for frontal images.

Table 5.6 summarizes the results of the 7 subjects for which the system correctly recognized 3 out of the 6 expressions and below. The table details the system’s response for each subject corresponding to each actual prototypic expression image. For convenience, cases that were correctly classified have been indicated with a “-”. As an arbitrary example, the fourth column of the table indicates that for images of the subjects for the expression “Anger”, the system incorrectly perceived “Disgust” for Subjects 1, 3 and 20, and “Sadness” for Subject 39, but correctly recognized “Anger” for Subjects 11, 30 and 33. The expressions (columns) in the table have been ordered according to the number of incorrect classifications, in descending order. Thus, the expression “Sadness” appears first because it was the most problematic among these cases. The Subjects (rows) have been similarly ordered. Thus, Subject 39 has been included first since this subject achieved the lowest classification rate.

TABLE 5.6: System response for subjects with 3 out of 6 expression recognition and below.

Subject	Sadness	Fear	Anger	Disgust	Surprise	Happiness
39	Anger	Disgust	Sadness	Anger	Disgust	-
33	Surprise	Anger	-	Surprise	-	Anger
1	Anger	-	Disgust	-	Fear	-
3	Anger	Disgust	Disgust	-	-	-
11	Anger	Surprise	-	-	-	Fear
20	Anger	Surprise	Disgust	-	-	-
30	Fear	-	-	Anger	Anger	-

Analyzing the table, it is noted that for these subjects, the misclassification were mostly concentrated in “Sadness“ and “Fear”, closely followed by “Anger”. It is noted that “Anger” is mostly confused with “Disgust”. It was initially expected that this was due

to similarity between the expressions, as explained previously. Focusing on expression “Sadness”, it is noted that this expression is confused with “Anger” in most cases. Analyzing the images for these subjects revealed that “Sadness” was expressed similarly to “Anger” in two such cases and expressed as a neutral expression in two cases. Figure 5.7 illustrates examples of such cases. A similar analysis of the expression “Fear” revealed that in most cases it was expressed as the neutral expression. Figure 5.8 illustrates examples of such cases.

This analysis revealed that, in addition to a similarity of expressions in some cases, inaccuracies in the data in the form of expressions closely resembling the neutral expression was also a source of error. The SVM was not trained on the neutral expression. Therefore, given an image of the neutral expression, an incorrect best-effort classification into one of the six classes would be carried out by the SVM. As such, the misclassification in many of these cases is attributed to the inaccurate manner in which the subjects performed these expressions, rather than inaccuracy of the system.

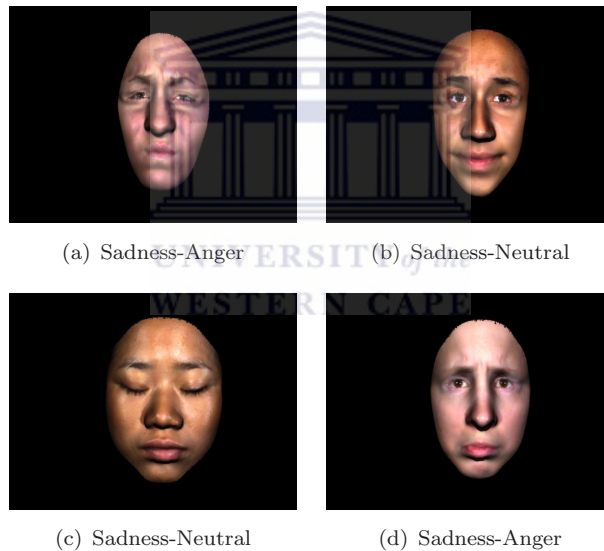


FIGURE 5.7: Examples of cases in which “Sadness” was expressed similarly to “Anger” or a neutral expression.

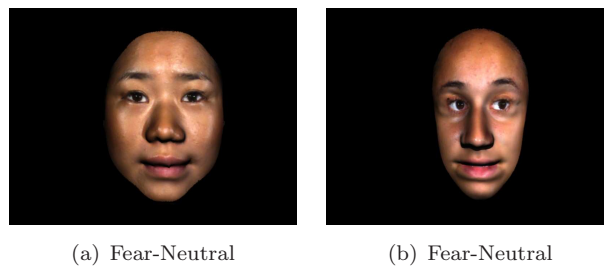


FIGURE 5.8: Examples of cases in which “Fear” was expressed similarly to the neutral expression.

Based on this finding, a deeper analysis of the data set was carried out by two independent assessors – Assessor 1 and Assessor 2 – to determine the number of images in the data set that resemble the neutral expression. Assessor 1 was shown frontal images of the first 20 test subjects – 150 images – and Assessor 2 was shown the remaining 20 test subjects – 150 images – from the BU-3DFE database. It should be noted that, as previously mentioned, the frontal facial model was used to produce the rotated image. The facial images in both cases are exactly the same. Therefore, this analysis was only carried out for frontal images. Table 5.7 summarizes the number of images per expression that were deemed to resemble the neutral expression. For reference, the full results of the neutral count for the 40 subjects for each expression are provided in Table A.3 in Appendix A.

TABLE 5.7: The total number of misclassified cases for each expression and the number of images of each expression that resemble the neutral expression.

Expression	Misclassified (40)	Neutral (40)
Anger	7	7
Disgust	15	13
Fear	15	9
Happiness	5	0
Sadness	12	10
Surprise	4	0

This result was surprising. The results indicate that in many cases – 39 cases – the facial expressions were performed similar to the neutral expression. The neutral expression accounts for 67% of the total number of misclassified expressions. It is noted that “Happiness” and “Surprise” were not affected at all by the neutral expression. However, all other expressions were heavily affected. In the case of “Anger”, all the misclassification are accounted for by the neutral expression. “Disgust”, “Sadness” and “Fear” have the highest number of samples that resembled the neutral expression, with the majority of misclassification for these expressions accounted for by the neutral expression. It is for this reason that these three expressions were the three lowest performing expressions.

It is important to note that, in spite of this fact, the proposed FER strategy was able to achieve high accuracies. This indicates that the strategy is highly robust. The results have been stated as is. However, for investigative purposes, those samples that resembled the neutral expression were removed to obtain an indication of the true recognition accuracy of the FER approach. Figure 5.9 visually depicts the results obtained by removing the affected samples.

The results indicate that all expressions have an accuracy of 85% and above. The two lowest performing expressions, “Disgust” and “Fear” increased to 95% and 85%,

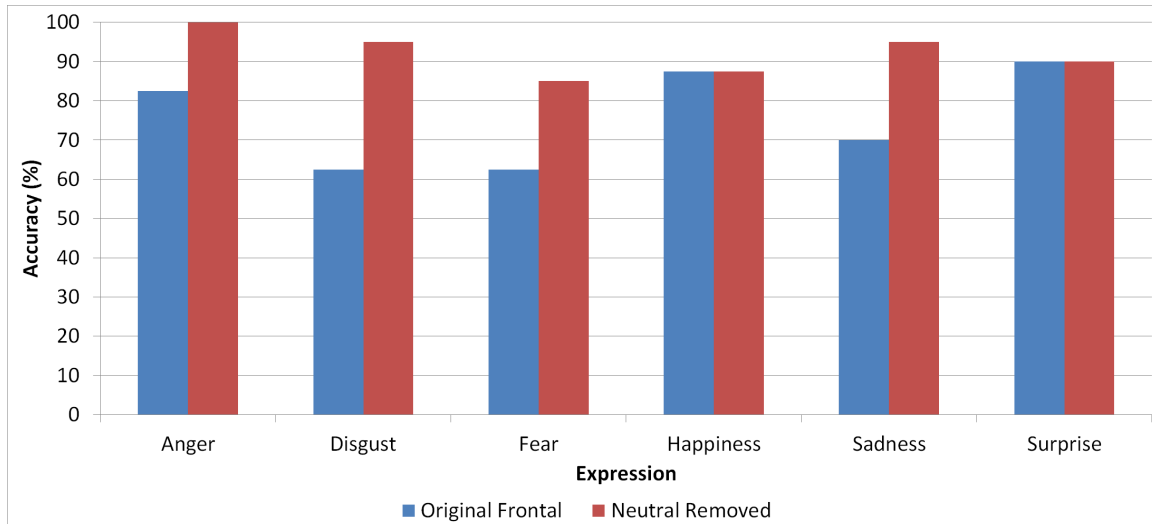


FIGURE 5.9: FER accuracy per expression for original frontal results and for the neutral cases removed.

respectively. “Anger” increased to 100% accuracy. These results further indicate that the FER strategy is highly robust and accurate. The results also indicate that the feature vector and SVM optimization procedures were highly successful.

The accuracy of this FER approach using a different locally collected data set was also determined. These results were published as part of this research. The paper is provided in [80] for reference. The data set consisted of 20 subjects – students of the University of the Western Cape – of varied skin tone performing each of the six prototypic expressions once. An image was captured from the frontal view and the rotated view. Each SVM was trained on images of 10 subjects and tested on images of the remaining 10 subjects. The frontal results on this data set were 85% across all expressions. Since this data set was manually collected, it was ensured that no instances of samples resembling the neutral expression were present. Hence, the results are higher. Once again, these results indicate that the FER system is robust and accurate.

5.3.3 Rotated Facial Images

This section discusses the experimental procedure carried out for rotated facial images and provides an analysis of the results obtained for each experiment.

5.3.3.1 Experimental Procedure

The rotated facial images from the BU-3DFE data set produced from the frontal images used in the previous section were used as input to the system. These images contain

faces rotated to 60° . In each case, the output of the system was analyzed using the criterion for accurately recognizing a facial expression.

5.3.3.2 Results and Analysis

For reference the full set of results obtained for the 40 subjects is provided in Table A.4 in Appendix A. The table provides the system response for each rotated image of each subject used as input. Table 5.8 is a confusion matrix summarizing the results after applying the FER strategy to rotated facial images. Table 5.9 summarizes the results as a percentage per expression and includes the frontal results for comparison.

TABLE 5.8: Results for rotated facial images.

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	25	4	5	2	4	0
Disgust	2	21	8	7	1	1
Fear	2	0	20	14	4	0
Happiness	0	1	1	38	0	0
Sadness	3	0	3	0	34	0
Surprise	1	0	5	1	1	32

TABLE 5.9: Comparison of average FER accuracy using frontal and rotated faces.

Expression	Frontal (%)	Rotated (%)
Anger	82	62
Disgust	62	52
Fear	62	50
Happiness	87	95
Sadness	70	85
Surprise	90	80
Average (%)	75	70

The system achieved an average recognition accuracy of 70% for rotated facial images ranging from 50% to 95%. This is lower than the average accuracy obtained for frontal images. It is noted that the system registered the highest accuracy of 95% for “Happiness”. It should also be noted that the system achieved a high average recognition accuracy of 80% and above for three of the expressions. Furthermore, the system registered an average accuracy of no less than 50% for all expressions. These results are highly encouraging considering that the images are rotated to an extreme angle of 60° . In terms of responding to the research question, it is clear that the proposed system can recognize whole expressions at a high accuracy in the presence of rotations of the face for non-occluded images.

It should be noted that, when the face is rotated to 60° , the most dominant feature area becomes the cheek region and sides of the face. Any features expressed in the nose, mouth or eye region become less pronounced. Therefore, any expressions that are primarily expressed using these regions are expected to register a reduction in FER accuracy, such as “Anger”, “Disgust”, “Fear” and “Surprise”. Conversely, any expressions primarily expressed in the cheek region or sides of the face are expected to register an increase in FER accuracy such as “Happiness” and “Sadness”.

The lowest accuracies, but by no means low, were again registered for the two expressions as in the frontal case: 50% for “Fear” and 52% for “Disgust”. Both of these accuracies were lower than the results for the corresponding frontal images. As in the frontal case, “Fear” was once again confused with “Happiness” in the majority of cases. 9 out of the 20 cases that were misclassified are attributed to the presence of “Fear” samples that resembled the neutral expression. The misclassification of the remaining 11 samples is attributed to: classification errors by the SVM; the fact that the expression is primarily expressed in the mouth region which is less visible and pronounced at 60° ; and the fact that the expression may appear similar to “Happiness”, as was the case for the frontal images.

It is also noted that the system incorrectly classified “Disgust” as “Fear” in the majority of cases. This is similarly attributed to the large number of cases – 13 cases – in which images labelled as “Disgust” resembled the neutral expression. In such cases, the system attempted an incorrect best-effort classification into one of the six classes, in this case mostly “Fear” and “Happiness”. The remaining 6 errors are attributed to classification errors by the SVM and the fact that the expression is primarily expressed in the nose and mouth regions, both of which are less visible and pronounced at 60° .

It was initially surprising to note that the recognition accuracy of the system increased for “Happiness” and “Sadness”. However, considering these expressions are primarily expressed in the cheek regions which are more pronounced and exposed in rotated images, it is expected that the FER accuracy would increase. One surprising aspect of this result is that, in spite of 10 samples of “Sadness” resembling the neutral expression, the expression obtained 34 out of 40 correct FER accuracy. This is surprising but may be attributed to the fact that, with a greater part of the cheek exposed, subtle features associated with “Sadness” may have become more pronounced in images that, on whole, may appear to resemble the neutral expression. This requires further investigation in future.

Moore and Bowden provide only visual results for the accuracy of the LBP^{ms} and LBP^{u2} operators at 60° . The visual results indicate an accuracy of between 60% and 70% for both operators. The results of the proposed system indicate a higher accuracy

of 70%. Once again, this result is only indicative since Moore and Bowden trained and tested on the entire data set. In future, the testing can be extended to the entire data set to perform a direct comparison. At this stage, it is only possible to state that this may indicate a greater effectiveness of the combined $LBP_{8,2}^{u2}$ operator over the two individual operators LBP^{ms} and LBP^{u2} for FER.

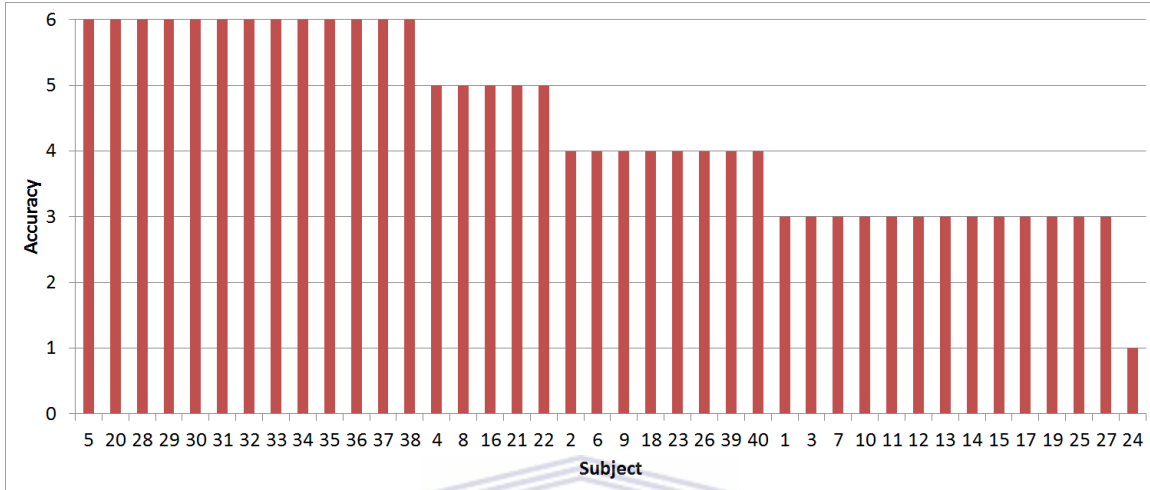


FIGURE 5.10: FER accuracy per subject for rotated images.

Figure 5.10 graphically summarizes the FER accuracy per subject for rotated images. For convenience, the results in the figure are once again sorted in descending order. For reference, the full set of results is provided in Table A.5 in Appendix A. An analysis of these results indicate that for 26 out of the 40 subjects – 65% of the subjects – the system correctly recognized at least 4 out of the 6 expressions. Furthermore, for 39 subjects – 97% of the subjects – the system correctly recognized at least 3 out of the 6 expressions. For only one subject, the system recognized less than 3 out of the 6 expressions. However, it should be noted that for this subject, the recognition rate of the system was 1 out of 6. For no subject did the system register 0 recognition. This results indicates that the system is highly robust to variations in test subjects at a rotated angle as well.

Figure 5.11 is a combination of the per-subject accuracy graphs of the frontal and rotated cases for comparison.

The graph illustrates that the accuracies for frontal and rotated cases are distributed randomly across test subjects. In order to investigate this further, the Pearson’s product-moment coefficient was computed between the two data sets to determine the correlation between the sets. The result was a value of $\rho = -0.098$ which indicates that the sets are poorly correlated. This indicates that the FER accuracy is independent of and invariant to test subjects.

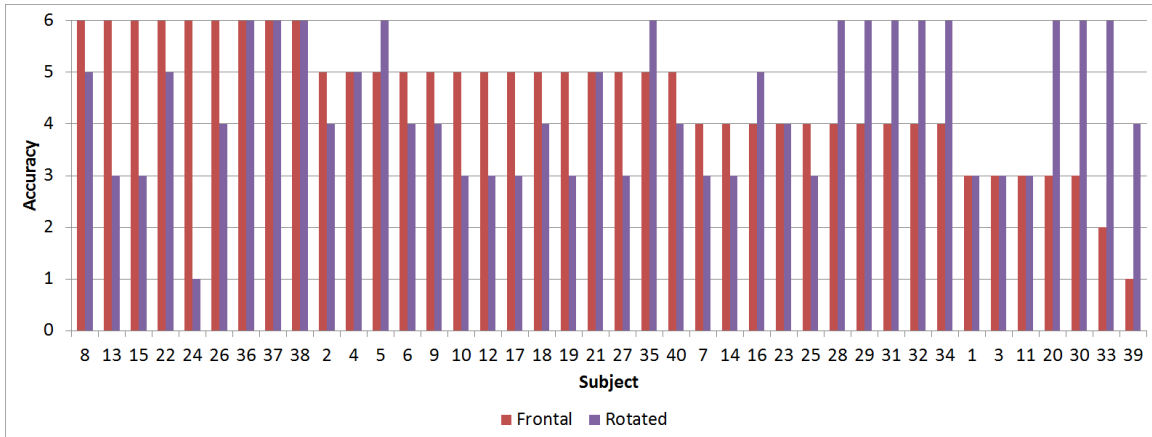


FIGURE 5.11: FER accuracy per subject for rotated images.

Table 5.10 summarizes the system response for Subject 24 for which the system correctly recognized 1 out of the 6 expressions.

TABLE 5.10: System response for Subject 24.

Expression	System Response
Anger	Sadness
Disgust	Sadness
Fear	Sadness
Happiness	Fear
Sadness	-
Surprise	Sadness

It is surprising to note that this subject achieved 100% correct recognition for the frontal case. The only correct classification was for “Sadness” which is represented as a “-”. Analyzing the table, it is noted that all but one of the system responses were “Sadness”. Figure 5.12 depicts the frontal and rotated images for this subject.

Analyzing the images, it is only possible to conclude that, in the frontal case, more features were exposed causing the SVM to achieve a better classification result than in the rotated case.

Once again, the results have been stated as is. However, for investigative purposes, those samples that resembled the neutral expression were removed to obtain an indication of the true recognition accuracy of the FER approach for rotated images. Figure 5.13 visually depicts the results obtained by removing the affected samples.

The results indicate that all expressions have an accuracy of 72% and above. The two lowest performing expressions, “Disgust” and “Fear” increased to 85% and 72%, respectively. “Anger” increased to 80% accuracy. These results further indicate that

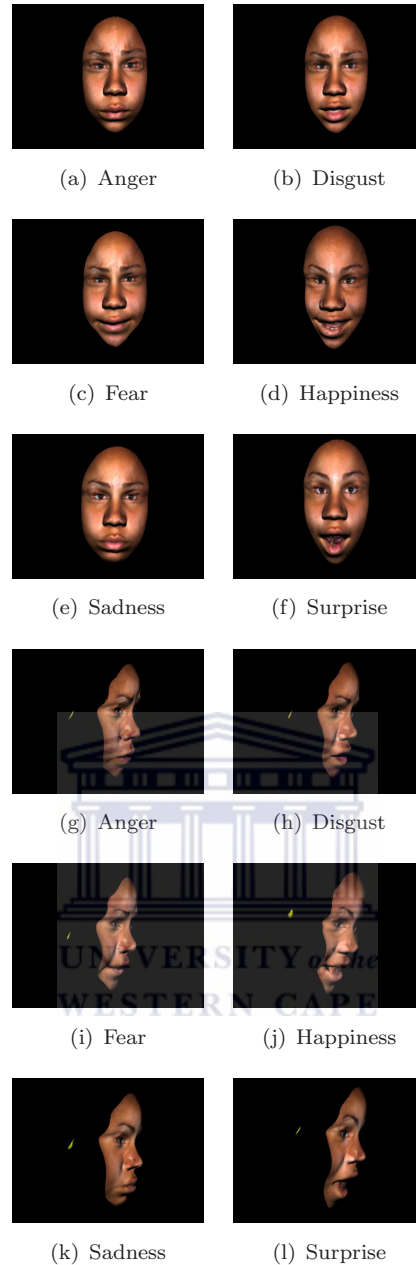


FIGURE 5.12: Frontal and rotated images for Subject 24.

the FER strategy is highly robust and accurate in the presence of rotations. The results also indicate that the feature vector and SVM optimization procedures were highly successful.

Once again, the accuracy of the FER approach on rotated images using a different locally collected data set was also determined and provided in the paper in [80] for reference. The rotated results on this data set were 80% across all expressions. Once again, since this data set was manually collected, it was ensured that no instances of samples resembling the neutral expression were present. Hence, the results are higher. These results indicate that the FER system is robust and accurate.



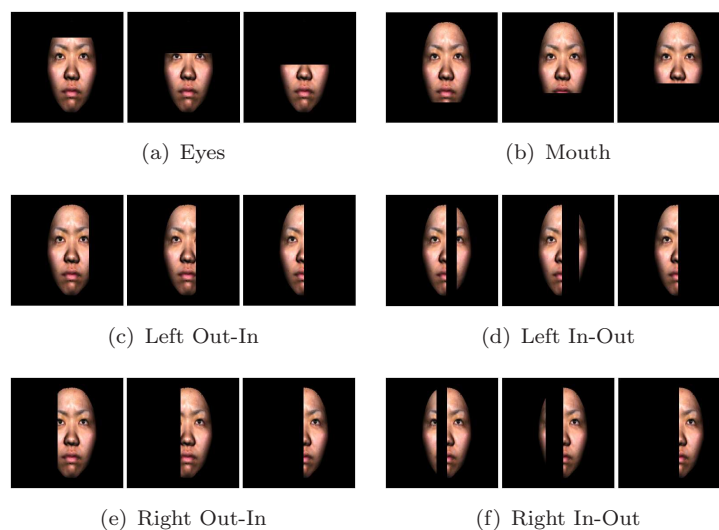
FIGURE 5.13: FER accuracy per expression for rotated images.

5.3.4 Frontal Occluded Facial Images

This section discusses the experimental procedure carried out for frontal occluded facial images and provides an analysis of the results obtained for each experiment.

5.3.4.1 Experimental Procedure

Figure 5.14 illustrates the simulation of the different levels of occlusion in the different regions.

FIGURE 5.14: Simulated partial occlusion for frontal images for each region at $(1/3)$, $(2/3)$ and full occlusion.

For frontal occluded images, partial occlusions of the eyes, mouth and left and right sides of the face were simulated by overlaying black pixels onto these regions similar to Kotsia *et al.* As an addition to Kotsia *et al.*'s work, three levels of occlusion were simulated at each region ranging from (1/3), (2/3) and full occlusion, which has not been done previously.

Additionally, for occlusions of the left and right sides of the face, occlusions were simulated from the outside to the inside as well as the inside to the outside of the face to compare the effects of occlusion in these two cases. According to literature, this has not been done previously either. For ease of reference, occlusion from the outside to the inside is henceforth referred to as "Out-In" and occlusion from inside to the outside is referred to as "In-Out". For example, occlusion of the left side Out-In means occlusion of the left side of the face from the outside to the inside of the face.

This was carried out in order to compare the effect of occlusion of features in the centre of the face (by occluding Out-In) and occlusion of features on the outside of the face (by occluding In-Out). It was expected that occlusion of features closer to the centre of the face should have a higher effect on FER accuracy.

This resulted in a total of 18 experiments, across 6 types of occlusion (mouth, eyes, left/right sides of the face inwards, left/right sides of the face outwards) and 3 levels of occlusion in each case. In each case, the output of the system was analyzed using the criterion for accurately recognizing a facial expression.

5.3.4.2 Results and Analysis

Table 5.11 summarizes all the different types and levels of frontal occlusion and provides the average accuracies across all expressions for the 18 experiments.

Analyzing the table, it can be seen that, even in the presence of full occlusion of various regions of the face, the proposed FER strategy is still able to recognize facial expressions at a high accuracy. This accuracy ranges from 70% for full occlusion of the left side of the face to a minimum of 58% for full occlusion of the mouth. In terms of responding to the research question, this clearly demonstrates that, even in the presence of severe partial occlusions of frontal facial images, the proposed FER approach can yield a high FER accuracy.

The average accuracy at full and no occlusion of each region is highlighted in bold text. Analyzing the table reveals that in the majority of cases – 88 out of 108 cases – occlusion resulted in a reduction in FER accuracy along with expectation. In a small number (20 of 108) of cases – about 18% of cases – the FER accuracy appears to slightly increase

TABLE 5.11: Results for each region and level of occlusion for frontal images.

		Anger(40)	Disgust(40)	Fear(40)	Happiness(40)	Sadness(40)	Surprise(40)	Average(%)
Eyes	No occlusion	33	25	25	35	28	36	75
	(1/3)	30	26	23	36	28	35	74
	(2/3)	35	27	24	34	27	32	74
	full	28	26	17	33	26	32	67
Mouth	No occlusion	33	25	25	35	28	36	75
	(1/3)	29	24	25	36	30	35	74
	(2/3)	20	26	22	38	24	32	67
	full	26	21	21	31	13	29	58
Left Out-In	No occlusion	33	25	25	35	28	36	75
	(1/3)	30	26	24	33	27	36	73
	(2/3)	35	25	23	32	23	36	72
	full	24	30	18	31	31	35	70
Left In-Out	No occlusion	33	25	25	35	28	36	75
	(1/3)	33	24	25	35	26	34	73
	(2/3)	29	24	21	34	27	33	70
	full	24	30	18	31	31	35	70
Right Out-In	No occlusion	33	25	25	35	28	36	75
	(1/3)	30	28	23	35	26	34	73
	(2/3)	23	24	26	35	28	30	69
	full	28	23	18	31	23	31	64
Right In-Out	No occlusion	33	25	25	35	28	36	75
	(1/3)	30	27	24	34	25	37	73
	(2/3)	29	19	23	37	28	36	71
	full	28	23	18	31	23	31	64

at varied levels of occlusion. Examples are: “Disgust” at all three levels of occlusion of the eye region and at full occlusion of the left side Out-In; “Sadness” at full occlusion of the left side Out-In and (1/3) occlusion of the mouth; and “Anger” at (2/3) occlusion of the eyes and (2/3) occlusion of the left side Out-In. The effect appears to be scattered at random across various expressions, facial regions and levels of occlusion. It may only be observed that “Disgust” benefited the most from this effect.

The result in these select cases is contrary to expectation since it is expected that any level of occlusion of any region should result in a reduction in salient features, therefore resulting in a reduction in FER accuracy or, at the very best, a sustained FER accuracy.

However, if it is taken into consideration that occlusion of parts of the face may result, in some cases, in a greater emphasis of regions rich in salient features and elimination of regions that are not rich in or are completely void of salient features, the observation of slight increases in FER accuracy in some cases may be expected. This indicates that a larger quantity of information is not necessarily always more conducive to achieving a higher FER accuracy.

This is analogous to the result of the region and resolution size optimization experiments in which a larger amount of information – a larger resolution size – did not necessarily yield a higher FER accuracy. In fact, the largest resolution size $R = (60 \times 60)$ yielded the lowest cross-validation accuracy for both frontal and rotated images despite providing the largest amount of information. A similar observation was made by Moore and Bowden. Moore and Bowden observed that the use of a large resolution size $R = (80 \times 110)$ did not result in the highest cross-validation accuracy in most cases, as explained in Chapter 2.

In addition, if the occlusion is viewed as a source of noise, Sheikh registered a similarly strange finding [99]. Sheikh investigated the effects of various types of noise on the FER accuracy using Gabor filters and SVMs for frontal images. Contrary to expectation, he found that the presence of a large amount of Gaussian and Poisson noise resulted in a higher FER accuracy than in facial images without any noise. He concluded that further investigation is required. As such, the occlusion that may be viewed as a source of noise may potentially contribute towards a higher FER accuracy. However, this result requires further investigation.

Focusing on the average accuracy over all six expressions, however, reveals that the result of occluding each region progressively from no occlusion to full occlusion results in an approximately continuous reduction in FER accuracy for each region, along with expectation. This illustrates that a global view of the results is along with expectation. This is depicted graphically in Figure 5.15. The first bar in each region is the average recognition accuracy for no occlusion of frontal facial images, a value of 75% obtained in a prior experiment.

Analysis of Figure 5.15 clearly demonstrates that occlusion of the mouth has the greatest effect on the FER accuracy than any other fully occluded region of the face. This finding coincides with Kotsia *et al.*'s [61] research. Furthermore, progressively occluding this region results in the most rapid deterioration in FER accuracy. This suggests that the mouth region plays the most pivotal role in the recognition of facial expressions. This also coincides with the large body of research that has consistently shown that the focus of the eye-gaze of Deaf signers within a conversation is concentrated on the mouth region [16, 78, 79].

The next region of importance, after the mouth, appears to be the right side of the face, both Out-In and In-Out. Occlusion of this area results in a much greater impact, and at a much more rapid rate on the FER accuracy, than the left side of the face, both Out-In and In-Out, as well as the eyes.

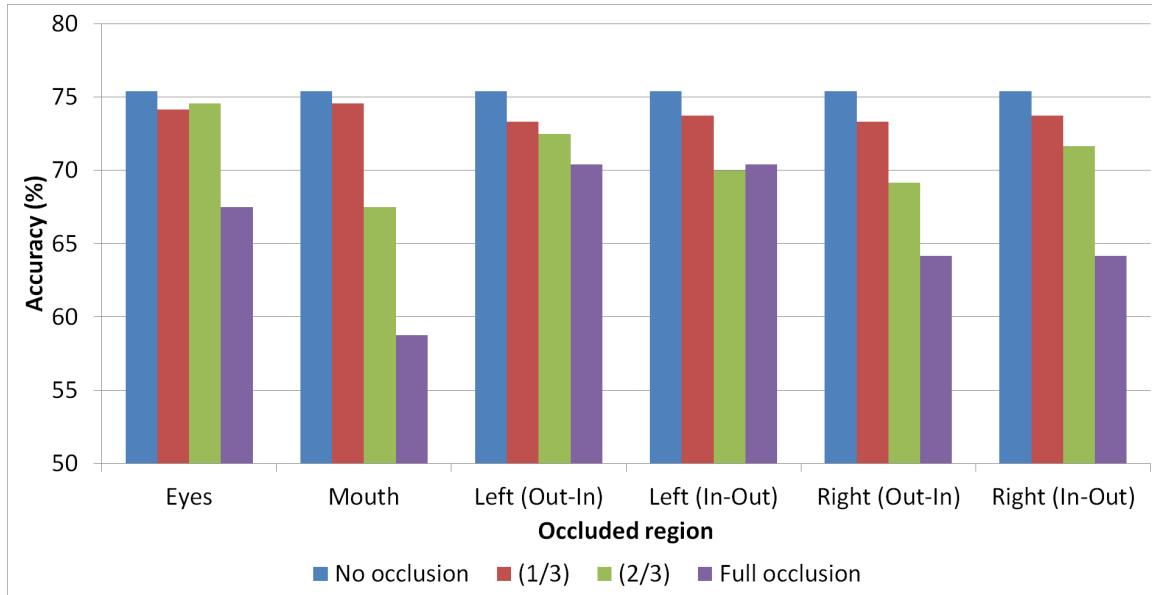


FIGURE 5.15: Average accuracy across each progressive level of occlusion across all expressions.

Finally, occlusion of the eye region appears to have a greater impact on the FER accuracy than occlusion of the left side of the face, both Out-In and In-Out, but only at full occlusion. At (1/3) and (2/3) occlusion of the eyes, there appears to be only a slight, if any, effect on the FER accuracy. This is not the case for occlusion of the left side of the face in which the FER accuracy appears to deteriorate more or less continuously as occlusion increases.

Comparing the effect of occlusion for Out-In and In-Out for both sides – comparing Left Out-In with Left In-Out and comparing Right Out-In and Right In-Out – of the face reveals that the effect of both types of occlusion – Out-In and In-Out – appear to be very similar. This indicates that, on average across all six expressions, the features on the outer parts of the face are as important as those in the centre of the face.

With regards to the effect of occlusion on the left and right sides of the face, the fact that occlusion of the right side and left side of the face appear to be different is an interesting finding. Since it is known that the six prototypic facial expressions are ideally symmetric, this finding may be attributed to the manner in which the subjects performed the six facial expressions. The expressions may be slightly more pronounced by the subjects in the right side of the face than in the left side of the face.

A similar but opposite result was obtained in the separate study carried out by the researcher provided in [80] on a different data set. In this case, it was found that occlusion of the left – not right – side of the face appeared to have a greater effect on the FER accuracy than the right side of the face and was attributed to the manner in which the subjects in that data set performed the expression i.e. with greater emphasis

on the left side of the face. This result indicates that, in order to be able to sustain a high accuracy over a large number of varied subjects, features from both sides of the face may be necessary to register a high FER accuracy to cater for cases of greater emphasis on either side of the face.

Figure 5.16 summarizes the average FER accuracy per expression at full occlusion of the left and right sides of the face. Analyzing this figure reveals that the greater emphasis on the right rather than the left side of the face by test subjects is mostly concentrated in three of the six expressions: “Disgust”, “Sadness” and “Surprise”. The indication that these expressions were more expressed on the right side than the left side of the face is obtained from the fact that occluding the right side resulting in a greater decrease in FER accuracy than occluding the left side. In the expressions “Fear” and “Happiness”, it appears that the expressions were performed symmetrically on average. Thus, occluding either side of the face resulted in an equal reduction in FER accuracy. For only “Anger”, it appears that test subjects performed the expression with a greater emphasis on the left side of the face, as evidenced by the fact that occluding the left side of the face resulted in a greater deterioration in FER accuracy than occluding the right side of the face.

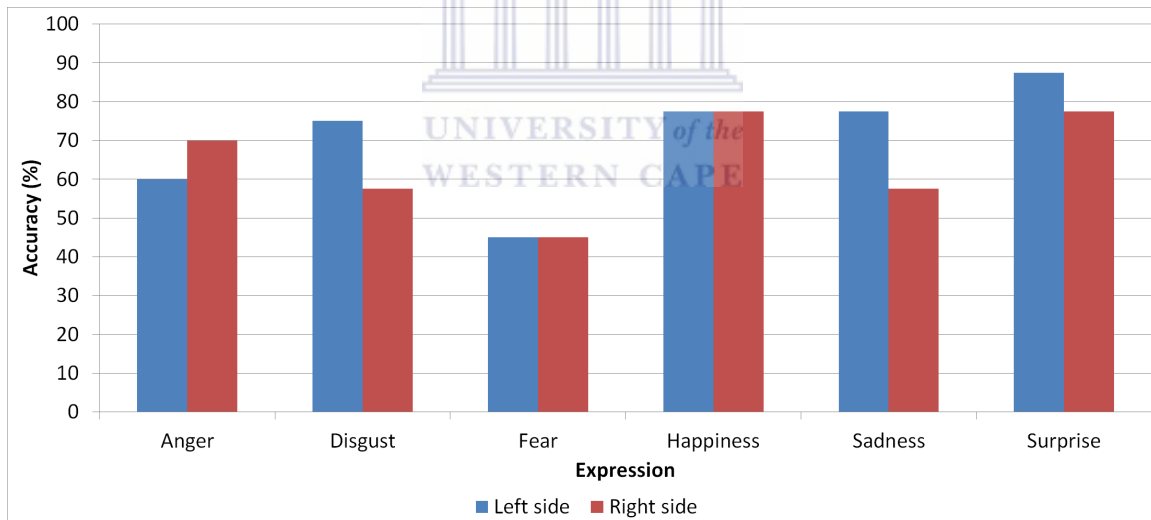


FIGURE 5.16: FER accuracy per expression for frontal images fully occluded on the left and right sides.

Figure 5.17 summarizes the average FER accuracy for each test subject at full occlusion across all regions. For reference, the full set of results is provided in Table A.6 in Appendix A.

An analysis of the graph shows that the system achieves higher than 80% accuracy for 7 of the 40 subjects, higher than 60% accuracy for 25 of the subjects – 62% of the subjects – and higher than 50% accuracy for 34 of the subjects – 85% of the subjects. For only 6 of the subjects does the system achieve lower than 50% accuracy. It is important to

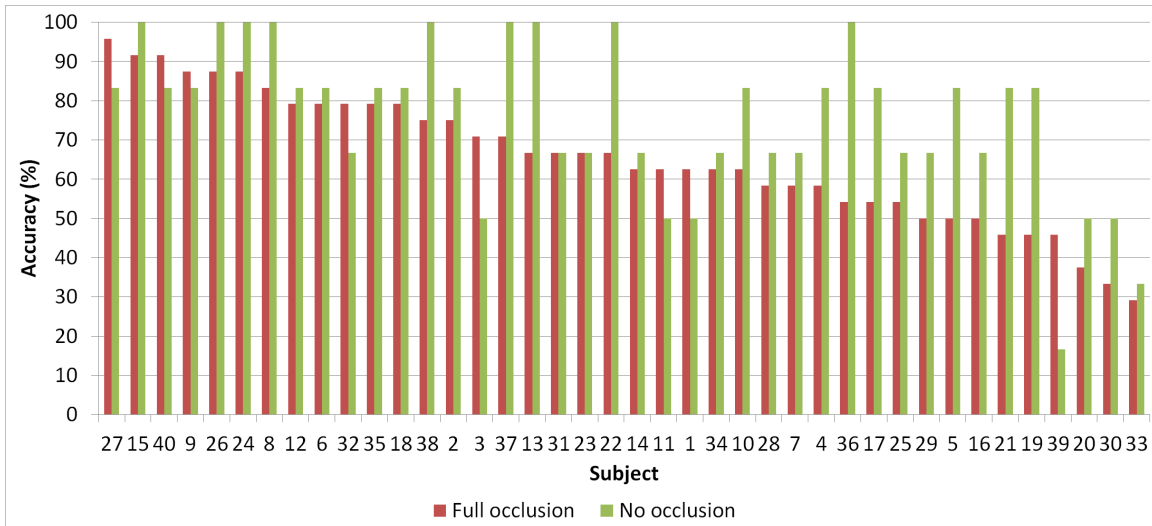


FIGURE 5.17: FER accuracy per subject for unoccluded and fully occluded frontal images.

note, however, that for no subject does the system achieve 0 recognition. The lowest accuracy is 29%.

The accuracy of the FER approach on frontal occluded images using a different locally collected data set was also determined and provided in the paper in [80] for reference. The average FER accuracy across all subjects at full occlusion for this data set ranged from 73% for full occlusion of the right side of the face to 45% for full occlusion of the mouth region. Once again, it was confirmed that occluding the mouth results in the greatest reduction in FER accuracy and this accuracy deteriorates at the most rapid rate as the level of occlusion progresses.

This data set appears, in general, to be more affected by occlusions than the BU-3DFE data set. Further investigation in this regard is required.

5.3.5 Rotated Occluded Facial Images

This section discusses the experimental procedure carried out for rotated occluded facial images and provides an analysis of the results obtained for each experiment.

5.3.5.1 Experimental Procedure

For rotated occluded images, partial occlusion of the eyes, mouth and the region between the eyes and the mouth were simulated by overlaying black pixels onto these regions of the rotated images, as illustrated in Figure 5.18. It was not possible to occlude the side of the face in the same manner as in the frontal images as this would result in a

complete occlusion of the entire face. As such, the region between the eyes and mouth was occluded instead. For ease of reference, the region between the eyes and mouth is henceforth referred to as the “middle region”. For example, (1/3) occlusion of the middle region means occluding the region between the eyes and mouth by (1/3).

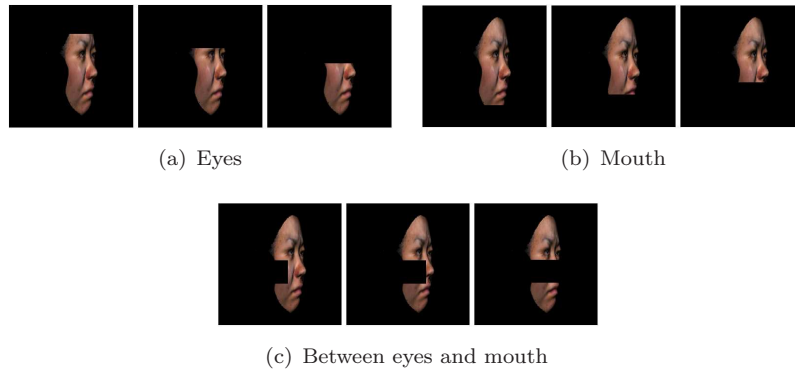


FIGURE 5.18: Simulated partially occluded rotated facial images at (1/3), (2/3) and full occlusion.

Once again, three levels of occlusion were simulated ranging from (1/3), (2/3) and full occlusion. This resulted in a total of 9 experiments, across 3 types of occlusion (mouth, eyes, middle region) and 3 levels of occlusion in each case. In each case, the output of the system was analyzed using the criterion for accurately recognizing a facial expression.

5.3.5.2 Results and Analysis

Table 5.12 summarizes the results of the experiments for all the different types and levels of rotated occlusion and provides the average accuracies across all expressions for the 9 experiments in the extreme right column. The average accuracy at full and no occlusion of each region is highlighted in bold text.

TABLE 5.12: Results for each region and level of occlusion for rotated images.

		Anger(40)	Disgust(40)	Fear(40)	Happiness(40)	Sadness(40)	Surprise(40)	Average(%)
Eyes	No occlusion	25	21	20	38	34	32	70
	(1/3)	31	14	17	27	15	32	56
	(2/3)	32	15	15	29	16	32	57
	full	30	18	14	27	17	28	55
Mouth	No occlusion	25	21	20	38	34	32	70
	(1/3)	28	11	16	25	16	33	53
	(2/3)	30	12	18	26	15	34	56
	full	28	10	17	17	14	34	50
Middle region	No occlusion	25	21	20	38	34	32	70
	(1/3)	31	9	18	23	15	31	52
	(2/3)	32	9	20	25	15	32	55
	full	30	11	18	25	17	31	55

Analyzing the table, it can be seen that, even in the presence of full occlusion of various regions of the face and rotation to an extreme angle of 60° , the proposed FER strategy is still able to recognize facial expressions at a high accuracy. This accuracy ranges from 55% for full occlusion of the eyes and middle region to a minimum of 50% for full occlusion of the mouth. In terms of responding to the research question, this clearly demonstrates that, even in the presence of severe partial occlusions of images of the face rotated to an extreme angle of 60° , the proposed FER approach can yield a high FER accuracy.

Analyzing the table reveals that in the majority of cases – 42 out of 54 cases – occlusion resulted in a reduction in FER accuracy along with expectation. As in the frontal occluded case, in a small number (12 of 54) of cases – about 22% of cases – the FER accuracy appears to slightly increase at varied levels of occlusion. In this case, only the following cases were registered: “Anger” at all three levels of occlusion for all facial regions; and “Surprise” at all levels of occlusion of the mouth region. This is different to the frontal occluded case in which the effect appeared to be scattered over a variety of expressions, regions and levels of occlusion.

As in the case of frontal occluded images, it may once again be taken into account that certain parts of the face contain noise – they are not rich in or completely void of salient features – and occluding these regions, in some cases, may result in a greater emphasis of regions that are rich in salient features. The fact that the effect manifests differently in the frontal occluded and rotated occluded images can be attributed to the intrinsic difference in features in the two cases. It has previously been explained that the recognition of facial expressions using LBP features is angle-specific. The resolution and region sizes of the frontal and rotated images are different. This implies that the classification model is different. A further detailed investigation is required in this regard.

The average accuracy at each level of occlusion for each region over all six expressions is depicted graphically in Figure 5.19. Analyzing the figure, it is observed that the result of occluding each region progressively from no occlusion to full occlusion is also different to the frontal occluded case. In this case, the application of (1/3) occlusion has a much more sudden and pronounced effect on the FER accuracy in all regions, but the FER accuracy appears to remain approximately constant with the subsequent application of (2/3) and full occlusion. In this case, it appears that, on average, (1/3), (2/3) and full occlusion appear to have an approximately equal effect on the FER accuracy and a more pronounced effect than in the frontal occluded case.

Further analyzing Figure 5.19, it is clear that for all regions and levels of occlusion, the system consistently recognizes facial expressions with accuracies ranging between 50% and 60%. Furthermore, it is very important to note that even at full occlusion

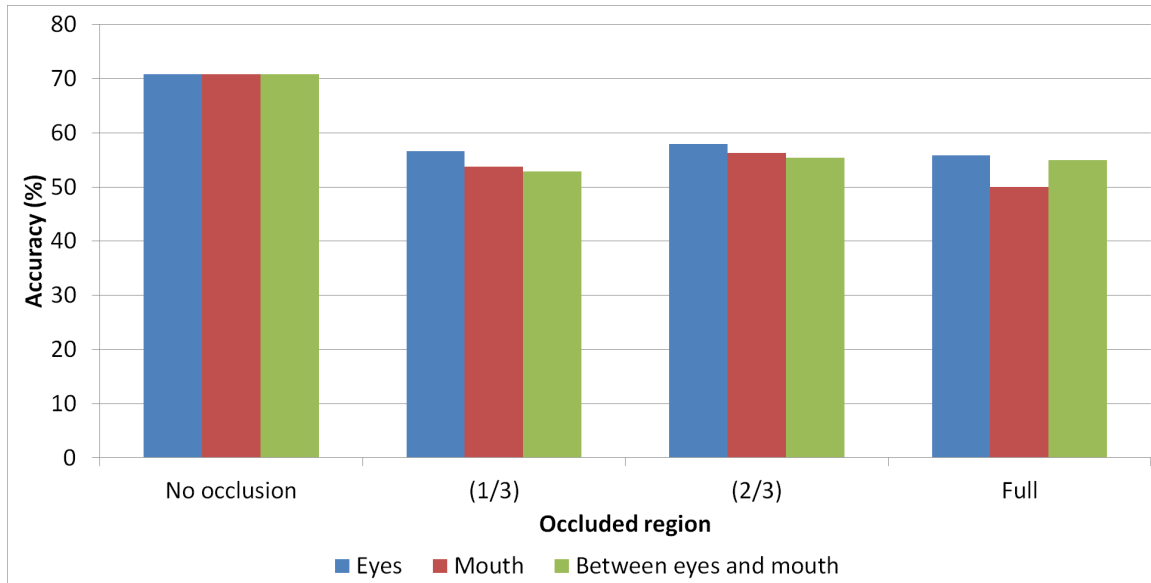


FIGURE 5.19: Average accuracy across all expressions progressively occluding each region of rotated images.

of all rotated facial regions, the system is still able to recognize facial expressions with accuracies of 50% and above. This is a very encouraging result since the images are rotated and the facial regions are fully occluded. This again illustrates that the proposed FER strategy is highly successful.

Similar to frontal, rotated and frontal occluded images, it is clear once again that for rotated occluded images the FER accuracy is mostly affected by occlusion of the mouth region. In this case, the middle region appears to be marginally more affected by occlusions than the eye region.

Figure 5.20 summarizes the average FER accuracy for each test subject rotated to 60° at full occlusion across all regions. For comparison, the same results for the frontal occluded case are also provided. For reference, the full set of results is provided in Table A.7 in Appendix A.

An analysis of the graph shows that the system achieves higher than 80% accuracy for 3 of the 40 subjects, higher than 60% accuracy for 17 of the subjects – 42% of the subjects – and an accuracy of 50% and higher for 26 of the subjects – 65% of the subjects. For 14 of the subjects, the system achieves lower than 50% accuracy. It is clear that the FER accuracy under these conditions is lower than under frontal, rotated and frontal occluded conditions. This result is as per expectation since the images are, both, rotated to an extreme angle of 60° and under full occlusion of a region. It is very encouraging to note, however, that under the most extreme conditions, for no subject does the system achieve 0 recognition. The lowest accuracy in this case is 22%. This clearly demonstrates that,

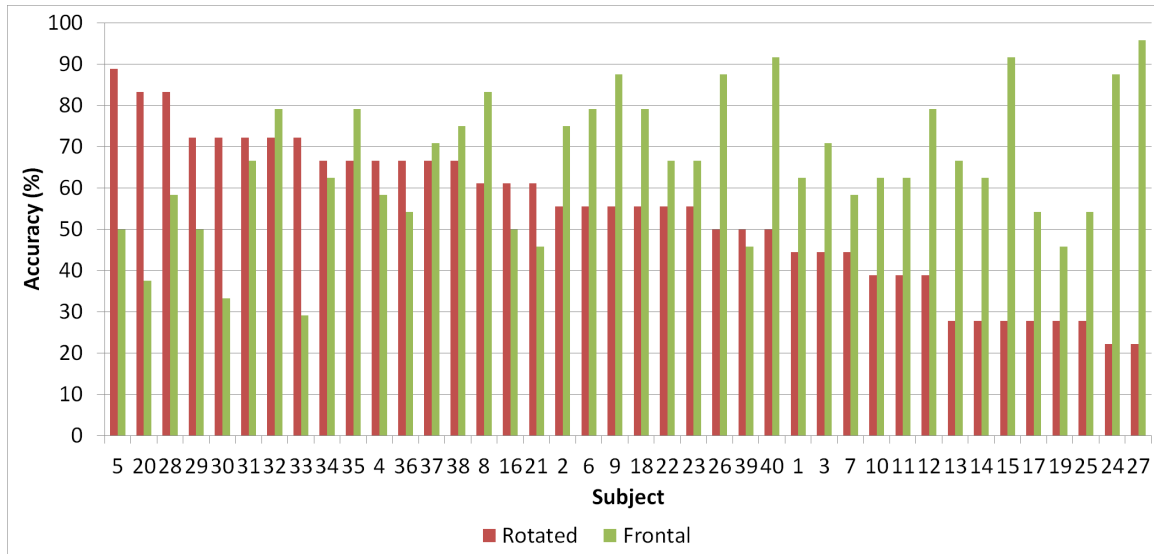


FIGURE 5.20: FER accuracy per subject for frontal and rotated images fully occluded for each region.

under the most extreme conditions, the proposed system is still robust to variations in test subjects.

A comparison of the frontal occluded per-subject results and the rotated occluded results illustrates, once again, that the accuracies for frontal and rotated cases are distributed randomly across test subjects. The Pearson’s product-moment coefficient was computed between the two data sets to determine the correlation between the sets. The result was a value of $\rho = -0.2140$ which indicates that the sets are poorly correlated. This indicates that the FER accuracy is independent of and invariant to test subjects under occlusions.

5.4 Summary and Conclusions

This chapter discussed the assessment of the proposed FER system. Various experiments were carried out in order to answer the two research questions: “Can the proposed face segmentation strategy accurately segment the face in facial images with varied skin tone, in the presence of rotations and on a complex background?”; and “Can whole facial expressions be recognized at a high accuracy using the LBP operator in the presence of rotations and partial occlusions of the face?”

In order to answer the first research question, an assessment of the face segmentation procedure was carried out. Two data sets were used as input to the proposed face segmentation procedure: a locally collected data set of 5 subjects with random skin tones on complex backgrounds and 50 subjects from the BU-3DFE database. For the locally collected data set, the face segmentation approach registered an accuracy of 100%

for all subjects except Subject 3, who registered an accuracy of 85%. It was found that an object to the right of this particular subject was incorrectly perceived as the face due to the similarity of the histograms of the subject's face and the detected object. The face segmentation procedure was then modified to only search for the face in a location not greater than 20% around the detected face in the initial frame. This resulted in a perfect face segmentation accuracy of 100% for all subjects. For the BU-3DFE data set, the system consistently achieved a perfect 100% accuracy across all subjects.

Therefore, in response to the first question, the face segmentation strategy can achieve a perfect face segmentation accuracy of 100% in facial images with varied skin tone, in the presence of rotations and on a complex background.

The second experiment carried out aimed at optimizing the resolution and region size for frontal and rotated images. This experiment went hand-in-hand with the optimization of the SVM. Using the cross-validation accuracy as a metric, different C and γ values as well as resolution and region values were evaluated. The combination of resolution and region size and the resulting C and γ values that achieved the highest cross-validation accuracy were deemed the optimum. This procedure was carried out for frontal and rotated images.

For frontal images, the optimum resolution size was 40×60 at an optimum region size of 8×10 . With these sizes, an optimum accuracy of 72.67% was obtained with $C = 0.5$ and $\gamma = 0.0078125$. For images rotated to 60° , the optimum resolution size was 40×50 at an optimum region size of 8×5 . With these sizes, an optimum accuracy of 69.67% was obtained with $C = 8.0$ and $\gamma = 3.0517578125 \times 10^{-05}$.

The resulting C and γ values were used to train two SVMs, one for frontal and one for rotated images. This experiment demonstrated that FER using the LBP operator is angle-specific as explained in Chapter 2.

In order to answer the second research question, experiments were carried out to assess the accuracy of the FER procedure. Separate experiments were carried out for the frontal, rotated, frontal occluded and rotated occluded images.

For frontal images, the FER system achieved a high average of 75% across all expressions and subjects. It was noted that the system is highly robust to variations in test subjects. An analysis of the causes of errors revealed the presence of a number of samples that resembled the neutral facial expression, but labelled as one of the six prototypic expressions. It was encouraging to observe that, in spite of the presence of such samples, the system was still able to recognize facial expressions at a high accuracy. It was shown that, with these samples removed, the average accuracy of the system could increase to 92%.

For rotated images, the average FER accuracy was 70% which was lower than the frontal case. It was, however, shown that the decrease was attributed to four of the six expressions, with two expressions – “Happiness” and “Sadness” – registering an increase in accuracy. It was stated that the frontal and rotated SVMs make use of different feature vectors but a further investigation in this regard was warranted.

Once again, a per-subject analysis demonstrated that the system is robust to variations in test subjects. For investigative purposes, an indication was provided as to the accuracy of the approach in the absence of the samples that resembled the neutral expression. It was shown that, in the absence of these samples, the FER accuracy could increase to 85%.

For frontal occluded images, the FER accuracy was shown to range between 70% for full occlusion of the left side of the face and 58% for full occlusion of the mouth. This result clearly demonstrates that the FER approach is robust to extreme occlusions of the face for frontal facial images. A surprising finding was made. It was found that, while in the majority of cases, occlusion led to a reduction in FER accuracy, in a small number of cases, occlusion caused a slight increase in FER accuracy. The effect was shown to be random across various regions of the face at varied levels of occlusion. “Disgust” was found to benefit from this effect the most. While further investigation was warranted, it was demonstrated that this may be caused by an occlusion of regions that are not rich in salient features, thereby providing a focus on salient-feature rich regions.

In general, the finding that the mouth is the most important region of the face in the literature was confirmed. The FER accuracy was shown to be most affected by occlusion of the mouth region, and affected at the most rapid rate. A per-subject analysis showed that the FER strategy is robust to variations in test subjects.

Finally, for rotated occluded images, the FER accuracy was shown to range from 55% for full occlusion of the eyes and middle region and 50% for full occlusion of the mouth. Once again, occlusion of the mouth region affected the FER accuracy the most. The FER accuracy for rotated occluded images was noted as being lower than frontal occluded images. However, even under these extreme conditions, the FER strategy achieved a highly encouraging result. A per-subject analysis once again confirmed that the system is robust to variations in test subjects.

As such, in response to the second research question, it is stated that the proposed FER system can recognize whole facial expressions at a high accuracy using the LBP operator in the presence of rotations and partial occlusions of the face.

A summary of the results obtained from the BU-3DFE database and the locally collected database are presented as accuracy ranges in Table 5.13

TABLE 5.13: Summary of the results obtained for the BU-3DFE database and the locally collected database.

Category	Range in Accuracy (%)	
	BU-3DFE database	Local database
Frontal	[62, 90]	[60, 90]
Rotated	[50, 95]	[70, 90]
Frontal Occluded	[58, 70]	[45, 75]
Rotated Occluded	[50, 55]	[-, -]

Chapter 6

Conclusion

In this research, several significant contributions towards the facial expression recognition (FER) component of the SASL system was made.

The experiments that were carried out in this research were ultimately aimed at answering the two research questions posed in Chapter 1:

1. “Can the proposed face segmentation strategy accurately segment the face in facial images with varied skin tone, in the presence of rotations and on a complex background?”
2. “Can whole facial expressions be recognized at a high accuracy using the LBP operator in the presence of rotations and partial occlusions of the face?”

The first contribution was the development of a highly accurate face segmentation procedure that is able to accurately segment frontal and rotated faces of various subjects containing varied skin tones in slightly different complex backgrounds. The videos contained subjects rotating their faces from left to right. The Viola-Jones face detector was not able to accurately detect the face in these conditions. This prompted the development of a face segmentation procedure using skin cues. The results for this procedure – a perfect face detection accuracy – revealed that the face segmentation procedure based on skin cues is promising in this respect. Prior to this research, the development of an automatic face segmentation procedure that accurately segments rotated faces in complex backgrounds has not been carried out by the SASL group. This is an important milestone for the SASL group since an automatic face segmentation procedure is a pre-requisite to any automatic facial expression recognition system.

The second contribution was the ability of the system to accurately recognize facial expressions in the presence of rotations and partial occlusions of the face. The use of the

accurate face segmentation procedure combined with the proposed Local Binary Pattern (LBP^c) operator used for feature extraction, is able to recognize frontal, rotated, frontal occluded and rotated occluded facial images with high average recognition accuracies. It should be noted that in addition to faces rotated to an extreme angle, the images also contained cases in which the eyes and mouth was fully occluded. Despite these images containing very limited facial information, the system is still able to recognize the six prototypic facial expressions with high average recognition accuracies. The results are extremely encouraging. Prior to this research, the investigation of rotated and partially occluded facial images on the FER accuracy has also not been carried out by the SASL group. Therefore, this is another important milestone for the group.

Furthermore, this research is novel since the effects of partially occluded facial images on the FER accuracy using LBPs were not investigated and the effects of rotated faces that are partially occluded were also not investigated in the literature, according to our knowledge.

6.1 Directions for Future Work

The directions for future work are provided in the following subsections.

6.1.1 Selecting a Suitable Database

It was found that the database used in this research contained a number of facial images that were labelled as a particular expression, but actually contained the neutral expression. Such a database will have a negative effect on the recognition accuracy of a system and the true accuracy of the system is thus hindered. Therefore, only experts should label images containing facial expressions.

Furthermore, no database containing facial expressions in SASL exists. Such databases are necessary to properly test for natural occlusion of the face by the hands.

6.1.2 Fully Automatic Systems

Fully automatic FER systems aid the development towards real-time FER systems. A fully automatic system is more practical in terms of its integration with other SASL systems towards the development of one fully-fledged SASL recognition system.

6.1.3 Comparing the Effects of Partial Occlusions and Rotations of the Face Using LBPs and Gabor Filters

According to the literature, researchers have not investigated the effects of rotations of the face on Gabor filters. Furthermore, researchers have also not investigated the effects of partial occlusions of the face on LBPs. A comparison of these two texture-based methods would reveal which method is more robust to the loss of facial information.

6.2 Concluding Remarks

Through the duration of this research, the researcher has gained a huge amount of experience. It is hoped that this research will serve as a base for other researchers pursuing the field of FER and add significant value towards the advancements of the SASL project.



Appendix A

Additional Test Results



TABLE A.1: System response for frontal images for the 40 subjects.

Subject	Anger(1)	Disgust(2)	Fear(3)	Happiness(4)	Sadness(5)	Surprise(6)
1	2	2	3	4	1	3
2	1	2	4	4	5	6
3	2	2	2	4	1	6
4	1	2	2	4	5	6
5	1	1	3	4	5	6
6	1	1	3	4	5	6
7	2	3	3	4	5	6
8	1	2	3	4	5	6
9	1	2	4	4	5	6
10	1	1	3	4	5	6
11	1	2	6	3	1	6
12	1	1	3	4	5	6
13	1	2	3	4	5	6
14	1	3	3	3	5	6
15	1	2	3	4	5	6
16	1	1	5	4	5	6
17	1	2	4	4	5	6
18	1	2	3	4	1	6
19	1	2	4	4	5	6
20	2	2	6	4	1	6
21	1	2	4	4	5	6
22	1	2	3	4	5	6
23	1	3	3	4	3	6
24	1	2	3	4	5	6
25	1	4	3	4	5	3
26	1	2	3	4	5	6
27	5	2	3	4	5	6
28	1	2	4	4	3	6
29	1	2	4	4	3	6
30	1	1	3	4	3	1
31	3	2	3	3	5	6
32	1	2	3	3	3	6
33	1	6	1	1	6	6
34	1	3	6	4	5	6
35	1	1	3	4	5	6
36	1	2	3	4	5	6
37	1	2	3	4	5	6
38	1	2	3	4	5	6
39	5	1	2	4	1	2
40	1	3	3	4	5	6

TABLE A.2: FER accuracy per subject for frontal images.

Subject	Correct (6)	Average (%)
1	3	50.00
2	5	83.33
3	3	50.00
4	5	83.33
5	5	83.33
6	5	83.33
7	4	66.67
8	6	100.00
9	5	83.33
10	5	83.33
11	3	50.00
12	5	83.33
13	6	100.00
14	4	66.67
15	6	100.00
16	4	66.67
17	5	83.33
18	5	83.33
19	5	83.33
20	3	50.00
21	5	83.33
22	6	100.00
23	4	66.67
24	6	100.00
25	4	66.67
26	6	100.00
27	5	83.33
28	4	66.67
29	4	66.67
30	3	50.00
31	4	66.67
32	4	66.67
33	2	33.33
34	4	66.67
35	5	83.33
36	6	100.00
37	6	100.00
38	6	100.00
39	1	16.67
40	5	83.33

TABLE A.3: Assessment of the frontal data set to determine the expressions that resemble the neutral expression (“1”) and those that do not (“0”).

Subject	Anger	Disgust	Fear	Happiness	Sadness	Surprise
1	0	0	1	0	1	0
2	1	0	0	0	1	0
3	0	1	0	0	1	0
4	0	0	0	0	0	0
5	1	1	1	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	1	0	0	1	0
9	1	1	0	0	0	0
10	1	0	1	0	0	0
11	0	0	1	0	1	0
12	0	1	0	0	0	0
13	0	1	0	0	0	0
14	0	0	0	0	1	0
15	0	0	0	0	1	0
16	0	0	0	0	0	0
17	0	0	0	0	0	0
18	0	1	1	0	1	0
19	0	1	0	0	0	0
20	0	0	0	0	0	0
21	0	1	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	0	0	0	0	0	0
25	0	0	0	0	0	0
26	1	0	0	0	0	0
27	0	1	0	0	0	0
28	0	0	0	0	0	0
29	0	0	0	0	0	0
30	0	0	0	0	1	0
31	0	1	1	0	0	0
32	0	0	0	0	0	0
33	1	1	1	0	1	0
34	0	1	0	0	0	0
35	0	0	0	0	0	0
36	1	0	0	0	0	0
37	0	0	0	0	0	0
38	0	0	0	0	0	0
39	0	0	1	0	0	0
40	0	0	1	0	0	0

TABLE A.4: System response for rotated images for the 40 subjects.

Subject	Anger(1)	Disgust(2)	Fear(3)	Happiness(4)	Sadness(5)	Surprise(6)
1	3	4	4	4	5	6
2	1	2	5	4	5	3
3	1	3	5	4	5	3
4	1	2	4	4	5	6
5	1	2	3	4	5	6
6	4	2	4	4	5	6
7	1	3	5	4	5	3
8	1	2	4	4	5	6
9	1	3	3	4	3	6
10	2	6	4	4	5	6
11	1	1	1	4	5	1
12	1	4	4	4	1	6
13	3	3	4	4	5	6
14	3	3	3	4	5	3
15	5	4	4	4	5	6
16	1	2	3	2	5	6
17	2	4	4	4	5	6
18	2	2	4	4	5	6
19	4	4	4	4	5	6
20	1	2	3	4	5	6
21	1	3	3	4	5	6
22	1	1	3	4	5	6
23	3	2	3	4	3	6
24	5	5	5	3	5	5
25	5	3	3	4	1	6
26	1	4	1	4	5	3
27	5	4	4	4	5	6
28	1	2	3	4	5	6
29	1	2	3	4	5	6
30	1	2	3	4	5	6
31	1	2	3	4	5	6
32	1	2	3	4	5	6
33	1	2	3	4	5	6
34	1	2	3	4	5	6
35	1	2	3	4	5	6
36	1	2	3	4	5	6
37	1	2	3	4	5	6
38	1	2	3	4	1	6
39	3	3	4	4	5	4
40	2	2	4	4	3	6

TABLE A.5: FER accuracy per subject for rotated images.

Subject	Correct (6)	Average (%)
1	3	50.00
2	4	66.67
3	3	50.00
4	5	83.33
5	6	100.00
6	4	66.67
7	3	50.00
8	5	83.33
9	4	66.67
10	3	50.00
11	3	50.00
12	3	50.00
13	3	50.00
14	3	50.00
15	3	50.00
16	5	83.33
17	3	50.00
18	4	66.67
19	3	50.00
20	6	100.00
21	5	83.33
22	5	83.33
23	4	66.67
24	1	16.67
25	3	50.00
26	4	66.67
27	3	50.00
28	6	100.00
29	6	100.00
30	6	100.00
31	6	100.00
32	6	100.00
33	6	100.00
34	6	100.00
35	6	100.00
36	6	100.00
37	6	100.00
38	6	100.00
39	4	66.67
40	4	66.67

TABLE A.6: Results for each region and level of occlusion for frontal images.

Subject	Eyes(6)	Mouth(6)	Left side(6)	Right side(6)
1	4	1	4	3
2	4	4	5	6
3	2	4	3	2
4	3	3	3	2
5	5	3	3	6
6	4	2	3	4
7	4	3	3	4
8	6	3	6	6
9	4	1	5	4
10	5	2	5	4
11	3	4	4	3
12	5	4	6	3
13	5	6	6	4
14	5	3	4	3
15	6	4	6	4
16	4	4	4	4
17	4	5	5	4
18	3	3	5	4
19	5	4	5	5
20	3	3	4	2
21	5	3	4	4
22	6	6	5	6
23	4	4	4	4
24	6	6	6	4
25	3	2	3	3
26	2	2	4	1
27	3	4	3	5
28	4	4	4	3
29	4	3	4	4
30	3	3	4	3
31	4	5	6	4
32	4	4	4	5
33	1	2	2	3
34	3	4	3	3
35	4	4	5	6
36	5	5	6	5
37	5	5	6	6
38	6	4	3	6
39	3	2	1	1
40	4	3	3	2

TABLE A.7: Results for each region and level of occlusion for rotated images.

Subject	Eyes(6)	Mouth(6)	Between eyes and mouth(6)
1	1	2	2
2	3	3	3
3	2	1	2
4	2	3	2
5	6	4	6
6	6	3	6
7	4	4	4
8	3	3	4
9	3	2	4
10	4	4	5
11	4	2	4
12	3	2	3
13	4	4	3
14	3	1	3
15	5	3	3
16	3	5	4
17	5	3	5
18	4	4	5
19	2	4	4
20	2	3	2
21	4	4	4
22	6	5	4
23	3	2	3
24	4	5	3
25	4	3	3
26	1	2	1
27	4	3	5
28	3	3	2
29	4	3	4
30	5	3	5
31	3	4	3
32	1	2	2
33	2	2	1
34	3	3	3
35	1	2	2
36	5	4	4
37	1	2	1
38	4	3	3
39	2	2	1
40	5	3	4

Bibliography

- [1] S. A Sirohey and A. Rosenfeld, “Eye detection in a face image using linear and nonlinear filters,” *Pattern recognition*, vol. 34, no. 7, pp. 1367–1391, 2001.
- [2] B. Abboud and F. Davoine, “Appearance factorization based facial expression recognition and synthesis,” in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4. IEEE, 2004, pp. 163–166.
- [3] I. Achmed, “Upper body pose recognition and estimation towards the translation of South African Sign Language,” Master’s thesis, University of the Western Cape, Computer Science, 2010.
- [4] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face recognition with local binary patterns,” in *Computer Vision-ECCV 2004*. Springer, 2004, pp. 469–481.
- [5] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face description with local binary patterns: Application to face recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 12, pp. 2037–2041, 2006.
- [6] Y. A. Alsultanny, “Color image segmentation to the rgb and hsi model based on region growing algorithm,” in *Proceedings of the 4th WSEAS international conference on Computer engineering and applications*. World Scientific and Engineering Academy and Society (WSEAS), 2010, pp. 63–68.
- [7] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, “Real time face detection and facial expression recognition: Development and applications to human computer interaction.” in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW’03. Conference on*, vol. 5. IEEE, 2003, pp. 53–53.
- [8] S. Birchfield, “Elliptical head tracking using intensity gradients and color histograms,” in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*. IEEE, 1998, pp. 232–237.
- [9] Blender Foundation, “Blender,” November 2013, [Online] Available at <http://www.blender.org/>.

- [10] D. S. Bolme, J. R. Beveridge, M. Teixeira, and B. A. Draper, "The csu face identification evaluation system: its purpose, features, and structure," in *Computer Vision Systems*. Springer, 2003, pp. 304–313.
- [11] F. Bourel, C. C. Chibelushi, and A. A. Low, "Recognition of facial expressions in the presence of occlusion." in *BMVC*. Citeseer, 2001, pp. 1–10.
- [12] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly, 2008.
- [13] D. Brown, "Faster upper body pose recognition and estimation using compute unified device architecture," Master's thesis, University of the Western Cape, Computer Science, 2012.
- [14] A. J. Calder and A. W. Young, "Understanding the recognition of facial identity and facial expression," *Nature Reviews Neuroscience*, vol. 6, no. 8, pp. 641–651, 2005.
- [15] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen, "Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps," *Neural Networks, IEEE Transactions on*, vol. 3, no. 5, pp. 698–713, 1992.
- [16] A. Cavender, R. E. Ladner, and E. A. Riskin, "Mobileasl: intelligibility of sign language video as constrained by mobile phone technology," in *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. ACM, 2006, pp. 71–78.
- [17] C. C. Chang and C. J. Lin, "LibSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [18] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold based analysis of facial expression," *Image and Vision Computing*, vol. 24, no. 6, pp. 605–614, 2006.
- [19] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade, "Feature-point tracking by optical flow discriminates subtle differences in facial expression," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 396–401.
- [20] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

- [21] I. Craw, D. Tock, and A. Bennett, "Finding face features," in *Computer Vision ECCV'92*. Springer, 1992, pp. 92–96.
- [22] N. Cristianini, "Support vector and kernel machines," *Tutorial at ICML*, 2001.
- [23] F. Dadgostar and A. Sarrafzadeh, "A fast real-time skin detector for video sequences," in *Image Analysis and Recognition*. Springer, 2005, pp. 804–811.
- [24] F. Dadgostar and A. Sarrafzadeh, "An adaptive real-time skin detector based on hue thresholding: A comparison on two motion tracking methods," *Pattern Recognition Letters*, vol. 27, no. 12, pp. 1342–1352, 2006.
- [25] D. Datcu and L. Rothkrantz, "Facial expression recognition in still pictures and videos using active appearance models: a comparison approach," in *Proceedings of the 2007 international conference on Computer systems and technologies*. ACM, 2007, p. 112.
- [26] H. Deldari and H. Sadoghi Yazdi, "Parallel implementation of eye detection algorithm on color facial images," *Australian Journal of Basic and Applied Sciences*, vol. 3, 2009.
- [27] H.-B. Deng, L.-W. Jin, L.-X. Zhen, and J.-C. Huang, "A new facial expression recognition method based on local gabor filter bank and pca plus lda," *International Journal of Information Technology*, vol. 11, no. 11, pp. 86–96, 2005.
- [28] S. A. P. Dept, "Algorithmic image matching (aim): Project analysis for santa ana police department," 1999.
- [29] I. Eibl-Eibesfeldt, *Human ethology*. Transaction Books, 2007.
- [30] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.
- [31] A. Elgammal, C. Muang, and D. Hu, "Skin detection-a short tutorial," *Encyclopedia of Biometrics*, 2009.
- [32] G. C. Feng and P. C. Yuen, "Multi-cues eye detection on gray intensity image," *Pattern recognition*, vol. 34, no. 5, pp. 1033–1046, 2001.
- [33] X. Feng, A. Hadid, and M. Pietikäinen, "A coarse-to-fine classification scheme for facial expression recognition," in *Image Analysis and Recognition*. Springer, 2004, pp. 668–675.

- [34] X. Feng, M. Pietikainen, and A. Hadid, "Facial expression recognition with local binary patterns and linear programming," *Pattern Recognition And Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 15, no. 2, p. 546, 2005.
- [35] R. Fisher, S. Perkins, A. Walker, and E. Wolfart, *Hypermedia image processing reference*. Wiley Chichester, UK, 1996.
- [36] M. M. Fleck, D. A. Forsyth, and C. Bregler, "Finding naked people," in *Computer Vision ECCV'96*. Springer, 1996, pp. 593–602.
- [37] M. Funk, K. Kuwabara, and M. J. Lyons, "Sonification of facial actions for musical expression," in *Proceedings of the 2005 conference on New interfaces for musical expression*. National University of Singapore, 2005, pp. 127–131.
- [38] T. Furness and Y. J. Lee, "Interaction control based on vision for AR interface of smart phone." *International Journal of Smart Home*, vol. 7, no. 4, 2013.
- [39] M. Ghaziasgar, "The use of mobile phones as service-delivery devices in a sign language machine translation system," Master's thesis, University of the Western Cape, Computer Science, 2010.
- [40] M. Glaser and W. Tucker, "Telecommunications bridging between deaf and hearing users in South Africa," in *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision and Hearing Impairments*, 2004.
- [41] H. P. Graf, T. Chen, E. Petajan, and E. Cosatto, "Locating faces and facial parts," in *Proc. First Intl Workshop Automatic Face and Gesture Recognition*, 1995, pp. 41–46.
- [42] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.
- [43] A. Gyaourova, C. Kamath, and S. Cheung, "Block matching for object tracking," *Lawrence livermore national laboratory*, 2003.
- [44] A. Hadid, M. Pietikainen, and T. Ahonen, "A discriminative feature space for detecting and recognizing faces," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–797.
- [45] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer vision and image understanding*, vol. 83, no. 3, pp. 236–274, 2001.

- [46] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1, pp. 185–203, 1981.
- [47] S. Howard, *Finger talk-South African sign language dictionary*. South Africa:Mondi, 2008.
- [48] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.
- [49] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 696–706, 2002.
- [50] J. Huang and H. Wechsler, "Eye detection using optimal wavelet packets and radial basis functions (rbfs)," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, no. 07, pp. 1009–1025, 1999.
- [51] W. Huang and R. Mariani, "Face detection and precise eyes location," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 4. IEEE, 2000, pp. 722–727.
- [52] M. Huenerfauth, "Generating american sign language classifier predicates for english-to-asl machine translation," Ph.D. dissertation, University of Pennsylvania, 2006.
- [53] P. T. Jackway and M. Deriche, "Scale-space properties of the multiscale morphological dilation-erosion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 1, pp. 38–51, 1996.
- [54] A. K. Jain, L. Hong, and S. Pankanti, "Mid-year population estimates," Statistics South Africa, Pretoria, South Africa, Tech. Rep. P0302, May 2013. [Online]. Available: <http://www.statssa.gov.za/Publications/statsdownload.asp?PPN=P0302>
- [55] L. Jordao, M. Perrone, J. P. Costeira, and J. Santos-Victor, "Active face and feature tracking," in *Image Analysis and Processing, 1999. Proceedings. International Conference on*. IEEE, 1999, pp. 572–576.
- [56] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.
- [57] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 46–53.

- [58] T. Kavzoglu and I. Colkesen, "A kernel functions analysis for support vector machines for land cover classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 11, no. 5, pp. 352–359, 2009.
- [59] S. Kawato and N. Tetsutani, "Detection and tracking of eyes for gaze-camera control," *Image and Vision Computing*, vol. 22, no. 12, pp. 1031–1038, 2004.
- [60] M. Kolsch and M. Turk, "Fast 2D hand tracking with flocks of features and multi-cue integration," in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*. IEEE, 2004, pp. 158–158.
- [61] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing*, vol. 26, no. 7, pp. 1052–1067, 2008.
- [62] J. Kovac, P. Peer, and F. Solina, *Human skin color clustering for face detection*. IEEE, 2003, vol. 2.
- [63] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *Computers, IEEE Transactions on*, vol. 42, no. 3, pp. 300–311, 1993.
- [64] M. P. Lewis, "Ethnologue: Languages of the world sixteenth edition," *Dallas, Tex.: SIL International. Online version: <http://www.ethnologue.com>*, 2009.
- [65] P. Li, "Hand shape estimation for South African Sign Language," Master's thesis, University of the Western Cape, Computer Science, 2010.
- [66] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE, 2002, pp. I–900.
- [67] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *Image processing, IEEE Transactions on*, vol. 11, no. 4, pp. 467–476, 2002.
- [68] W. F. Liu and Z. Wang, "Facial expression recognition based on fusion of multiple gabor features," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE, 2006, pp. 536–539.
- [69] A. Lotriet, "Sign language interpreting in south africa: Meeting the challenges," *Critical Link*, 2001.

- [70] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision." in *IJCAI*, vol. 81, 1981, pp. 674–679.
- [71] J. Luettin, N. A. Thacker, and S. W. Beet, "Visual speech recognition using active shape models and hidden markov models," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 817–820.
- [72] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces," *Stockholm, Sweden: Karolinska Institute*, 1998.
- [73] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 200–205.
- [74] A. Maruch, "Talking with the hearing-impaired," January 2010, [Online] Available at <http://www.deafsa.co.za/>.
- [75] S. J. McKenna, S. Gong, and Y. Raja, "Modelling facial colour and identity with gaussian mixtures," *Pattern recognition*, vol. 31, no. 12, pp. 1883–1892, 1998.
- [76] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Computer Vision ECCV 2008*. Springer, 2008, pp. 504–513.
- [77] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Computer Vision and Image Understanding*, vol. 115, no. 4, pp. 541–558, 2011.
- [78] L. Muir, I. Richardson, and S. Leaper, "Gaze tracking and its application to video coding for sign language," in *Picture Coding Symposium*, 2003, pp. 23–25.
- [79] L. J. Muir and I. E. Richardson, "Perception of sign language and its application to visual communications for deaf people," *Journal of Deaf Studies and Deaf Education*, vol. 10, no. 4, pp. 390–401, 2005.
- [80] D. Mushfieldt, M. Ghaziasgar, and J. Connan, "Robust facial expression recognition in the presence of rotation and partial occlusion," in *Proceedings of the 2013 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*. ACM, 2013, pp. 186–193.
- [81] N. Naidoo, "South African Sign Language recognition using feature vectors and hidden markov models," Master's thesis, University of the Western Cape, Computer Science, 2009.

- [82] K. Nallaperumal, S. Ravi, C. N. K. Babu, R. Selvakumar, A. L. Fred, C. Seldev, and S. Vinsley, "Skin detection using color pixel classification with application to face detection: a comparative study," in *Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on*, vol. 3. IEEE, 2007, pp. 436–441.
- [83] J. A. Nasiri, S. Khanchi, and H. R. Pourreza, "Eye detection algorithm on facial color images," in *Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on*. IEEE, 2008, pp. 344–349.
- [84] J. A. Nasiri, M. A. Moulavi, S. N. Gelyan, H. Deldari, H. S. Yazdi, and A. E. Shargh, "An efficient parallel eye detection algorithm on facial color images," in *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD'08. Ninth ACIS International Conference on*. IEEE, 2008, pp. 706–711.
- [85] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.
- [86] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.
- [87] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.
- [88] T. Otsuka and J. Ohya, "Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences," in *Image Processing, 1997. Proceedings., International Conference on*, vol. 2. IEEE, 1997, pp. 546–549.
- [89] T. Otsuka and J. Ohya, "Spotting segments displaying facial expression from image sequences using HMM," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 442–447.
- [90] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 84–91.
- [91] P.-T. Pham-Ngoc and Q.-L. Huynh, "Robust face detection under challenges of rotation, pose and occlusion," 2010.

- [92] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification." in *NIPS*, vol. 12, 1999, pp. 547–553.
- [93] C. Rajah, "Chereme-based recognition of isolated, dynamic gestures from South African Sign Language with Hidden Markov Models," Master's thesis, University of the Western Cape, Computer Science, 2006.
- [94] J. H. Relethford, "Human skin color diversity is highest in sub-saharan african populations," *Human Biology*, pp. 773–780, 2000.
- [95] R. Schweiger, P. Bayerl, and H. Neumann, "Neural architecture for temporal emotion classification," in *Affective Dialogue Systems*. Springer, 2004, pp. 49–52.
- [96] K. Seshadri and M. Savvides, "Robust modified active shape model for automatic facial landmark annotation of frontal faces," in *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*. IEEE, 2009, pp. 1–8.
- [97] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
- [98] C. Shan and T. Gritti, "Learning discriminative lbp-histogram bins for facial expression recognition." in *BMVC, 2008*, pp. 1–10.
- [99] M. Sheikh, "*Robust Recognition of Facial Expressions on Noise Degraded Facial Images*," Master's thesis, University of the Western Cape, Computer Science, 2011.
- [100] M. C. Shin, K. I. Chang, and L. V. Tsap, "Does colorspace transformation make any difference on skin detection?" in *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*. IEEE, 2002, pp. 275–279.
- [101] L. Sigal, S. Sclaroff, and V. Athitsos, "Estimation and prediction of evolving color distributions for skin segmentation under varying illumination," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2. IEEE, 2000, pp. 152–159.
- [102] W. Skarbek, A. Koschan, and Z. Veroffentlichung, "Colour image segmentation-a survey," 1994.
- [103] W. C. Stokoe, "Sign language structure: An outline of the visual communication systems of the american deaf," *Journal of deaf studies and deaf education*, vol. 10, no. 1, pp. 3–37, 2005.

- [104] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press-Elsevier, 2003.
- [105] Y.-I. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 97–115, 2001.
- [106] A. Tzotsos and D. Argialas, “Support vector machine classification for object-based image analysis,” in *Object-Based Image Analysis*. Springer, 2008, pp. 663–677.
- [107] B. Vadapalli, “Recognition of facial action units from video streams with recurrent neural networks : a new paradigm for facial expression recognition,” Ph.D. dissertation, University of Western Cape, 2012.
- [108] H. Van Kuilenburg, M. Wiering, and M. Den Uyl, “A model based method for automatic facial expression recognition,” in *Machine Learning: ECML 2005*. Springer, 2005, pp. 194–205.
- [109] V. Vezhnevets, V. Sazonov, and A. Andreeva, “A survey on pixel-based skin color detection techniques,” in *Proc. Graphicon*, vol. 3. Moscow, Russia, 2003, pp. 85–92.
- [110] L. Vezzano, “Icaam - inverse compositional active appearance models,” September 2013, [Online] Available at <http://www.mathworks.com>.
- [111] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.
- [112] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.
- [113] G. S. Wachtman, J. F. Cohn, J. M. VanSwearingen, and E. K. Manders, “Automated tracking of facial features in patients with facial neuromuscular dysfunction,” *Plastic and reconstructive surgery*, vol. 107, no. 5, pp. 1124–1133, 2001.
- [114] J. Whitehill, “Automatic real-time facial expression recognition for signed language translation,” Master’s thesis, University of the Western Cape, Computer Science, 2006.

- [115] J. Whitehill, "Automatic real-time facial expression recognition for signed language translation," Master's thesis, University of the Western Cape, Computer Science, 2006.
- [116] T. Wilhelm, H.-J. Böhme, and H.-M. Gross, "Classification of face images for gender, age, facial expression, and identity," in *Artificial Neural Networks: Biological Inspirations-ICANN 2005*. Springer, 2005, pp. 569–574.
- [117] L. Yi and J. Connan, "Kerntune: Self-tuning linux kernel performance using support vector machines," in *Proceedings of the 2007 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*. ACM, 2007, pp. 189–196.
- [118] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. IEEE, 2006, pp. 211–216.
- [119] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *Neural Networks, IEEE Transactions on*, vol. 17, no. 3, pp. 683–695, 2006.
- [120] B. D. Zarit, B. J. Super, and F. K. Quek, "Comparison of five color models in skin pixel classification," in *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999. Proceedings. International Workshop on*. IEEE, 1999, pp. 58–63.
- [121] C. Zor, "Facial expression recognition," Master's thesis, University of Surrey, Guildford, 2008.