

UNIVERSITY OF THE WESTERN CAPE

# Autonomous Facial Expression Recognition using the Facial Action Coding System



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

March 2016

# Declaration



I, Nathan de la Cruz, declare that this thesis “Autonomous Facial Expression Recognition using the Facial Action Coding System” 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: ..... *Ndelacruz* .....

Date: ..... 16 March 2016 .....

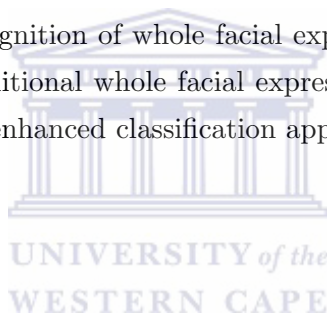
*“Imagination is more important than knowledge. For knowledge is limited to all we now know and understand, while imagination embraces the entire world, and all there ever will be to know and understand.”*

Albert Einstein



# Abstract

The South African Sign Language research group at the University of the Western Cape is in the process of creating a fully-fledged machine translation system to automatically translate between South African Sign Language and English. A major component of the system is the ability to accurately recognise facial expressions, which are used to convey emphasis, tone and mood within South African Sign Language sentences. Traditionally, facial expression recognition research has taken one of two paths: either recognising whole facial expressions of which there are six i.e. anger, disgust, fear, happiness, sadness, surprise, as well as the neutral expression; or recognising the fundamental components of facial expressions as defined by the Facial Action Coding System in the form of Action Units. Action Units are directly related to the motion of specific muscles in the face, combinations of which are used to form any facial expression. This research investigates enhanced recognition of whole facial expressions by means of a hybrid approach that combines traditional whole facial expression recognition with Action Unit recognition to achieve an enhanced classification approach.

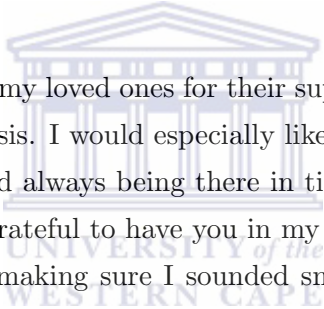


# Keywords

South African Sign Language, Facial Expression Recognition, Facial Action Coding System, Action Units, Whole Facial Expressions, Face Detection, Haar Features, Dense Optical Flow, Support Vector Machine.



# Acknowledgements



I would like to thank all of my loved ones for their support and encouragement throughout the duration of my thesis. I would especially like to thank Cassidy for her words of wisdom and motivation and always being there in times of stress. Your words inspired and uplifted me. I am so grateful to have you in my life. Also, thank you for reviewing and editing my work and making sure I sounded smarter than what I actually am. I would like to thank my mother for her prayers and her unwavering support. You would sacrifice the world for me and for that reason I am truly blessed.

I would like to thank my supervisor, Mehrdad Ghaziasgar for his guidance and patience with me. You have gone above and beyond what is expected of a supervisor. I thank you for all of your help in the completion of this research. To my co-supervisor James Connan, Thank you for your support and guidance on how to approach this research. And finally, to my colleagues Roland, Waleed and Kenzo, I thank you for always making it a pleasure to come to campus everyday. You all have made my experience at the University of the Western Cape a good one.

# Contents

<b>Declaration of Authorship</b>	<b>i</b>
<b>Abstract</b>	<b>iii</b>
<b>Keywords</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>x</b>
<b>Abbreviations</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Research Question . . . . .	4
1.3 Research Objectives . . . . .	4
1.4 Premises . . . . .	5
1.5 Methodology . . . . .	5
1.6 Thesis Outline . . . . .	7
<b>2 Related Work</b>	<b>8</b>
2.1 Action Unit Recognition Systems . . . . .	8
2.2 Whole Facial Expression Recognition Systems . . . . .	14
2.3 Hybrid Facial Expression Recognition Systems . . . . .	20
2.4 Summary and Conclusion . . . . .	26
<b>3 Image Processing Techniques for Facial Expression Recognition</b>	<b>28</b>
3.1 Face Detection and Segmentation . . . . .	28
3.1.1 Haar-Like Wavelet Feature Detection . . . . .	29
3.1.2 Integral Image . . . . .	30
3.1.3 AdaBoost Learning Algorithm . . . . .	31
3.1.4 Producing a Rejection Cascade of Weak Classifiers . . . . .	31



3.1.5	Analysis of the Viola-Jones Face Detection System . . . . .	32
3.2	Feature Extraction . . . . .	33
3.2.1	Dense Optical Flow Tracking . . . . .	33
3.2.2	Farneback Dense Flow Polynomial Expansion . . . . .	35
3.2.3	Estimation of Displacement . . . . .	36
3.3	Classification . . . . .	37
3.3.1	SVM Classification Process . . . . .	38
3.3.2	Adaptation of the SVM for Non-Linear Problem Classification . . . . .	41
3.3.3	Kernel Functions . . . . .	42
3.3.4	Multi-class SVM Classification Techniques . . . . .	42
3.3.4.1	One-Versus-All . . . . .	43
3.3.4.2	One-Versus-One . . . . .	43
3.3.4.3	Directed Acyclic Graph . . . . .	43
3.4	Summary . . . . .	44
<b>4</b>	<b>Design and Implementation of the Facial Expression Recognition System</b>	<b>45</b>
4.1	Face Detection and Segmentation . . . . .	45
4.1.1	Whole Face Segmentation . . . . .	47
4.1.2	Upper Face Segmentation . . . . .	48
4.1.3	Lower Face Segmentation . . . . .	48
4.2	Feature Extraction . . . . .	49
4.3	Classification . . . . .	51
4.3.1	WFE Method . . . . .	51
4.3.2	AU Method . . . . .	51
4.3.3	HybridWFEFirst Method . . . . .	56
4.3.4	HybridAUFIRST Method . . . . .	56
4.4	Training of Classifiers . . . . .	59
4.4.1	Training Dataset . . . . .	59
4.4.2	SVM Optimization Procedure . . . . .	60
4.4.3	Training of the AU Classifiers . . . . .	60
4.4.4	Training of the Whole Facial Expression Multi-Class Classifiers . . . . .	61
4.5	Summary . . . . .	63
<b>5</b>	<b>Design and Implementation of the Facial Expression Recognition System</b>	<b>64</b>
5.1	AU Recognition Accuracy Experiment . . . . .	65
5.1.1	Criterion for a Correctly Recognised AU . . . . .	65
5.1.2	Experimental Procedure . . . . .	65
5.1.3	Results and Analysis . . . . .	66
5.1.3.1	Overview of Results . . . . .	67
5.1.3.2	Global VS Local Segmentation for AU Recognition . . . . .	71
5.2	Facial Expression Recognition Accuracy Experiment . . . . .	73
5.2.1	Criterion for a Correctly Recognised Facial Expression . . . . .	73
5.2.2	Experimental Procedure . . . . .	73
5.2.3	Results and Analysis . . . . .	74
5.2.3.1	WFE Method Accuracy Results and Analysis . . . . .	74

5.2.3.2	AU Method Accuracy Results and Analysis . . . . .	77
5.2.3.3	HybridWFEFirst Method Accuracy Results and Analysis	79
5.2.3.4	HybridAUFfirst Method Accuracy Results and Analysis .	81
5.2.3.5	Comparison of Methods . . . . .	83
5.3	Summary and Conclusions . . . . .	87
<b>6</b>	<b>Conclusion</b>	<b>89</b>
6.1	Future Work . . . . .	90
6.2	Concluding Remarks . . . . .	91
<b>A</b>	<b>Additional Test Results</b>	<b>92</b>
	<b>Bibliography</b>	<b>97</b>



# List of Figures

1.1	The DSR methodology’s iterative cycle [69]. . . . .	6
2.1	Custom template consisting of 22 feature points used by Kapoor <i>et al.</i> [37].	9
2.2	The set of upper face AUs recognised by Kapoor <i>et al.</i> ’s system [37]. . . . .	9
2.3	System overview of Lien <i>et al.</i> ’s system [44]. . . . .	11
2.4	AUs recognised by Lien <i>et al.</i> ’s system and their descriptions [44]. . . . .	12
2.5	An example of the feature point tracking used by Lien <i>et al.</i> [44]. . . . .	12
2.6	An example of dense optical flow computation [44]. . . . .	13
2.7	Features that are manually placed on the image in the normalisation method used by Cohn <i>et al.</i> [11]. . . . .	14
2.8	The six basic emotional expressions [13]. . . . .	15
2.9	Active Appearance Models Used by Datcu and Rothkranz [13]. . . . .	15
2.10	Isolated frontal face obtained by Mushfieldt <i>et al.</i> ’s face segmentation procedure [49]. . . . .	17
2.11	Isolated side-view of the face obtained by Mushfieldt <i>et al.</i> ’s face segmentation procedure [49]. . . . .	18
2.12	The normalisation procedure used by Mushfieldt <i>et al.</i> to correct for misalignment of the face [49]. . . . .	18
2.13	Superimposed grid of feature points used by Schweiger <i>et al.</i> [62]. . . . .	19
2.14	Facial points inserted on the frontal-view of the face as used by Pantic <i>et al.</i> [55]. . . . .	21
2.15	Facial points inserted on the side view of the face, used by Pantic <i>et al.</i> [55]. . . . .	23
2.16	An example of the feature map used by Pantic <i>et al.</i> [55]. . . . .	23
2.17	Active contour models computed for the eyebrows, eyes and mouth by Pantic <i>et al.</i> [55]. . . . .	24
2.18	Manually initialised regions drawn on the face by Yacoob and Davis [74].	25
3.1	Haar-like features [68]. . . . .	29
3.2	A visual description of the integral image representation [68]. . . . .	30
3.3	An example of an integral image computed from a target image [68]. . . . .	31
3.4	Computation a haar-like feature using lookups from the integral image [68].	31
3.5	A rejection cascade of weak classifiers [68]. . . . .	32
3.6	Various examples of the Viola-Jones face detection algorithm in operation [68]. . . . .	32
3.7	Example frames of running and writing and their resultant dense optical motion flows [40]. . . . .	35
3.8	The dense flow displacement computation over two frames [24]. . . . .	37
3.9	A two-class classification problem [12]. . . . .	38

3.10	The optimal hyperplane that separates the two classes with a maximum margin [12]. . . . .	39
3.11	Directed Acyclic Graph of a 4-class problem [58]. . . . .	44
4.1	Processing overview of the proposed FER system. . . . .	46
4.2	The Viola-Jones algorithm detects the face. . . . .	47
4.3	The Viola-Jones algorithm detects the eye-pair. . . . .	47
4.4	Isolated face after the whole face segmentation procedure. . . . .	48
4.5	Isolated face after the upper face segmentation procedure. . . . .	48
4.6	Isolated face after the lower face segmentation procedure. . . . .	48
4.7	An example motion flow, showing its independent horizontal and vertical flows. . . . .	50
4.8	The six basic emotional expressions recognised. . . . .	52
4.9	An overview of the <i>WFE</i> method. . . . .	52
4.10	An overview of the <i>AU</i> method. . . . .	53
4.11	Depictions and descriptions of the 16 AUs in the upper and lower face. . .	54
4.12	An overview of the <i>HybridWFEFirst</i> method. . . . .	57
4.13	An overview of the <i>HybridAUFIRST</i> method. . . . .	58
5.1	Graphical depiction of the AU classifier recognition accuracy results for the global segmentation method (“GS”) and local segmentation method (“LS”). . . . .	68
5.2	Depictions and descriptions of the 16 AUs in the upper and lower face. . .	69
5.3	A graph depicting the average of the GS and LS accuracy for each AU (on the right vertical axis), sorted in ascending order of the number of training examples available for each AU (on the left vertical axis). . . . .	70
5.4	A graph depicting the average of the GS and LS accuracy for each AU (on the right vertical axis), sorted in ascending order of the number of training examples available for each AU (on the left vertical axis). . . . .	71
5.5	Recognition accuracy of the <i>WFE</i> method. . . . .	75
5.6	Recognition accuracy of the <i>AU</i> method. . . . .	78
5.7	Recognition accuracy of the <i>HybridWFEFirst</i> method. . . . .	80
5.8	Recognition accuracy of the <i>HybridAUFIRST</i> method. . . . .	82
5.9	A graphical depiction of the average accuracy of each method across all six emotions. . . . .	84
5.10	A graphical depiction of the accuracy of each FER method per expression. .	86

# List of Tables

2.1	AU recognition accuracy of Kapoor <i>et al.</i> 's system [37]. . . . .	10
2.2	Head gesture accuracy of Kapoor <i>et al.</i> 's system [37]. . . . .	10
2.3	Number of samples in the Cohn-Kanade dataset used by Datcu and Rothkranz [13]. . . . .	16
2.4	Confusion matrix of the static approach of Datcu and Rothkranz [13]. . . . .	16
2.5	Confusion matrix of the temporal approach of Datcu and Rothkranz [13]. . . . .	17
2.6	Frontal and rotated FER accuracy of Mushfieldt <i>et al.</i> 's system [49]. . . . .	18
2.7	Confusion matrix of Schweiger <i>et al.</i> 's classifier [62]. . . . .	20
2.8	Production rules used by Pantic <i>et al.</i> to infer the six basic emotional expressions from combinations of AUs [55]. . . . .	20
2.9	Facial features used by Pantic <i>et al.</i> to characterise frontal and rotated faces [55]. . . . .	22
2.10	Mouth and chin features used by Pantic <i>et al.</i> to characterise frontal faces only [55]. . . . .	22
2.11	Confusion matrix of Pantic <i>et al.</i> 's classifier [55]. . . . .	24
2.12	Yacooob and Davis' dictionary that describes the local directional motions in the mouth region, where $W$ denotes the rectangular window around the feature [74]. . . . .	25
3.1	Processing time (in ms) per frame of the three dense optical flow methods compared by Le Bek, averaged over 100 frames [40]. . . . .	35
4.1	Production rules used to infer the six basic emotional expressions using AUs [55]. . . . .	55
4.2	Intrinsic accuracy $V_i$ of, and the type of segmentation ("Segm. Type") used by, each final AU classifier. . . . .	55
4.3	Reference maximum value $M_{P_j}$ of the production rule $P_j$ of each expression $E_j$ . . . . .	56
4.4	Number of positive sequences available in the dataset for each AU and the number of sequences used to train each AU classifier. . . . .	61
4.5	Optimized parameter values for each AU classifier for global and local segmentation. . . . .	62
4.6	Number of sequences available in the dataset for each emotion and the number of sequences of each emotion used to train the multi-class classifiers. . . . .	62
4.7	Optimized parameter values for the multi-class classifier of the <i>WFE</i> and <i>HybridAUFIRST</i> method. . . . .	63
5.1	Number of positive sequences available in the dataset for each AU and the number of sequences used to test each AU classifier. . . . .	66

5.2	AU classifier recognition accuracy results for the global segmentation method (“GS”) and local segmentation method (“LS”). . . . .	67
5.3	The testing data used in the FER accuracy experimentation. . . . .	74
5.4	Facial expression recognition results of the <i>WFE</i> method. . . . .	74
5.5	Confusion matrix of the <i>WFE</i> method accuracy results. . . . .	76
5.6	Facial expression recognition results of the <i>AU</i> method. . . . .	77
5.7	Confusion matrix of the <i>AU</i> method accuracy results. . . . .	78
5.8	Facial expression recognition results of the <i>HybridWFEFirst</i> method. . . .	79
5.9	Confusion matrix of the <i>HybridWFEFirst</i> method accuracy results. . . . .	81
5.10	Facial expression recognition results of the <i>HybridAUFfirst</i> method. . . . .	82
5.11	Confusion matrix of the <i>HybridAUFfirst</i> method accuracy results. . . . .	83
5.12	Results of McNemar’s test for comparing the four FER methods. . . . .	85
A.1	AU classifier recognition accuracy results. . . . .	92
A.2	Facial expression recognition results of the four methods: AU, WFE, HybridWFEFirst and HybridAUFfirst. . . . .	93
A.3	Confusion matrix of the WFE method accuracy results. . . . .	94
A.4	Confusion matrix of the AU method accuracy results. . . . .	94
A.5	Confusion matrix of the HybridWFEFirst method accuracy results. . . . .	94
A.6	Confusion matrix of the HybridAUFfirst method accuracy results. . . . .	94
A.7	McNemar’s Test for the WFE and AU methods. . . . .	95
A.8	McNemar’s Test for the WFE and HybridWFEFirst methods. . . . .	95
A.9	McNemar’s Test for the WFE and HybridAUFfirst methods. . . . .	95
A.10	McNemar’s Test for the AU and HybridWFEFirst methods. . . . .	95
A.11	McNemar’s Test for the AU and HybridAUFfirst methods. . . . .	95
A.12	McNemar’s Test for the HybridWFEFirst and HybridAUFfirst methods. . .	96



# Abbreviations

<b>FACS</b>	<b>F</b> acial <b>A</b> ction <b>C</b> oding <b>S</b> ystem
<b>FER</b>	<b>F</b> acial <b>E</b> xpression <b>R</b> ecognition
<b>AU</b>	<b>A</b> ction <b>U</b> nit
<b>WFE</b>	<b>W</b> hole <b>F</b> acial <b>E</b> xpression
<b>SASL</b>	<b>S</b> outh <b>A</b> frican <b>S</b> ign <b>L</b> anguage
<b>DSR</b>	<b>D</b> esign <b>S</b> cience <b>R</b> esearch
<b>CUDA</b>	<b>C</b> ompute <b>U</b> nified <b>D</b> evice <b>A</b> rchitecture
<b>SVM</b>	<b>S</b> upport <b>V</b> ector <b>M</b> achines
<b>HMM</b>	<b>H</b> idden <b>M</b> arkov <b>M</b> odel
<b>LDA</b>	<b>L</b> inear <b>D</b> iscriminant <b>A</b> nalysis
<b>PCA</b>	<b>P</b> rincipal <b>C</b> omponent <b>A</b> nalysis
<b>AAM</b>	<b>A</b> ctive <b>A</b> ppearance <b>M</b> odel
<b>CCA</b>	<b>C</b> onected <b>C</b> omponents <b>A</b> nalysis
<b>ISFER</b>	<b>I</b> ntegrated <b>S</b> ystem for <b>F</b> acial <b>E</b> xpression <b>R</b> ecognition
<b>LBP</b>	<b>L</b> ocal <b>B</b> inary <b>P</b> atterns
<b>GS</b>	<b>G</b> lobal <b>S</b> egmentation
<b>LS</b>	<b>L</b> ocal <b>S</b> egmentation
<b>CPU</b>	<b>C</b> entral <b>P</b> rocessing <b>U</b> nit
<b>GPU</b>	<b>G</b> raphics <b>P</b> rocessing <b>U</b> nit
<b>GB</b>	<b>G</b> igabyte
<b>GHz</b>	<b>G</b> iga <b>h</b> ertz
<b>RAM</b>	<b>R</b> andom <b>A</b> ccess <b>M</b> emory
<b>IR</b>	<b>I</b> nfra - <b>R</b> ed
<b>BU-3DFE</b>	<b>B</b> inghamton <b>U</b> niversity <b>3D</b> <b>F</b> acial <b>E</b> xpression
<b>RBF</b>	<b>R</b> adial <b>B</b> asis <b>F</b> unction
<b>DAG</b>	<b>D</b> irected <b>A</b> cyclic <b>G</b> raph

---

<b>ROI</b>	<b>R</b> egion <b>O</b> f <b>I</b> nterest
<b>CK+</b>	extended <b>C</b> ohn - <b>K</b> anade



# Chapter 1

## Introduction

### 1.1 Background and Motivation

Facial expressions are universally indistinguishable; each are innate human traits that are commonly found all over the world, suggesting that they can be characterised and recognised. This idea is supported by the research of Ekman and Friesen which models muscle movements in the face and can be extended to characterise facial expressions [19, 20].

It was not always a known fact that facial expressions are universal and consistent across cultures. It was a fiercely contested subject as anthropologists and psychologists had been grappling with this question for decades. Darwin suggested that facial expressions are universally similar based on his theory of evolution [17]. However, the research community were not convinced as there was no general consensus. Ekman and Friesen conducted studies on subjects from eastern and western cultures in 1971, and they concluded that facial expressions were indeed similar across cultures [17]. Even though the results of this study were well accepted, Russell questioned the fact that facial expressions could be universally recognised and wrote a paper critiquing Ekman and Friesen's results in 1994 [60]. Later that year, Izard [35] and Ekman [16] responded to Russell's critique with strong evidence, and refuted the claims that Russell made. Since then, it has been an established fact that facial expressions can be recognised across cultures.

The model developed by Ekman and Friesen is known as the Facial Action Coding System (FACS) [18], where individual or combinations of distinct facial muscle movements are identified by Action Units (AUs) [15]. The FACS defines 44 unique AUs. Of these, 30 AUs are linked to the contraction of muscles in the face, made up of 12 muscles

situated in the upper region of the face and 18, situated in the lower region of the face. It was observed that over 7000 distinct AU combinations are possible [61]. Ekman found that a subset of AUs related to contractions in the face can be coded to describe six basic emotional expressions, namely: Happy, Sadness, Anger, Disgust, Fear and Surprise [19, 20]. These are also referred to as whole facial expressions (WFEs).

A substantial amount of research has been geared towards recognising AUs using computer vision. Applications of these vary from deception detection [43, 57], to emotion detection [22] and sign language recognition [49]. The research presented in this thesis is done in the context of sign language recognition and undertaken as a part of the South African Sign Language (SASL) project at the University of the Western Cape.

The SASL project involves the development of a real-time machine translation system that seamlessly translates between English and SASL [30]. It is a necessary part of this system to use computer vision to extract semantic information from a video of a deaf person communicating in SASL. The semantic information extracted from sign language video is characterized by five fundamental sign language parameters [30, 33, 49]: hand motion, hand orientation, hand location, hand shape, and facial expressions. The first four parameters are collectively referred to as manual gesture parameters.

The SASL project has carried out extensive research in recognising manual gesture parameters. Achmed [1, 2] developed a system that detects the location of the hands and motions of the arms. Brown [7, 8] extended Achmed's work to run on the Graphics Processing Unit (GPU) using the Compute Unified Device Architecture (CUDA) to enhance the processing speed of the existing system. Li [41, 42] developed a hand shape estimation system utilizing a 3D avatar to render the hand shapes. Foster [25, 26] used Li's feature extraction procedure and carried out an extensive comparison of machine learning techniques to determine the most accurate and appropriate technique with the approach.

On a separate but related front, Rajah [59] and Naidoo [50, 51] developed systems to recognise SASL gestures based only on the extracted hand motion parameter. Nel [53, 54] and Frieslaar [27, 28] greatly extended the gesture recognition capabilities of these two systems by combining two parameters towards gesture recognition. Nel extracted and used the hand shape and location, while Frieslaar used the hand shape and motion to characterize gestures.

The SASL project has also carried out research into facial expression recognition, and this will be mentioned shortly. Facial expressions communicate non-verbal cues and help to convey tone in conversation. They are also an important component of sign language communication as mood and tonality are conveyed by the face, and both can be

misconstrued without facial expressions. Research has found that a deaf individual's eye-gaze is concentrated mainly on the facial region when in a sign-language conversation, particularly around the mouth region [10, 48].

For this reason, the proposed research focuses on recognising facial expressions. The majority of facial expression recognition (FER) systems fall into one of two major classes which shall henceforth be referred to as the "traditional" approaches. The first class of systems aim to recognise sub-units of facial expressions, mainly by recognising individual AUs or combinations of AUs, without any form of collation to infer, or concern towards recognising, WFEs [23, 44, 64]. The second class of systems aim to recognise WFEs on a global scale, disregarding AUs or any other facial subunits altogether [47, 56].

A third class of systems also exists in the form of hybrid systems that first recognise smaller subunits of facial expressions such as AUs and subsequently use the recognised subunits as descriptors to recognise the six basic emotional expressions. One such system [55] does so by means of a set of production rules proposed by Ekman and Friesen [18]. These rules explicitly specify various AUs that are present while each of the six basic emotional expressions are performed. Generally, however, very little research has been conducted in this area, especially regarding using AUs to recognise WFEs. More importantly, no comparisons have been carried out to determine how such hybrid approaches may compare with traditional WFE recognition approaches.

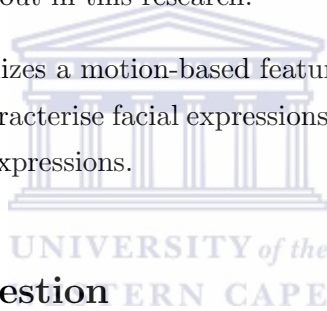
In terms of research into the recognition of facial expressions, the SASL project has mainly focused on the first class of systems in which AUs in the face are recognised using the FACS. Whitehill [70, 71] recognised muscle movements in the face characterised by AUs using Haar features and the AdaBoost algorithm for classification. Whitehill compared the effectiveness of using global versus local segmentation in his endeavour. Sheikh [63] utilised Support Vector Machines (SVMs) trained on Gabor-filter images to create an AU recognition system and analyse the effect of noise degraded images on the system. Vadapalli [66, 67] developed an AU recognition system using Gabor filters to be trained using two machine learning techniques, namely, recurrent neural networks and SVMs. All of these systems were found to be highly successful at detecting and recognising AUs.

Along with the above research, the SASL project has also conducted research into recognising WFEs corresponding to the second class of systems. Mushfieldt [49] created a FER system that recognised the six basic emotional expressions at different levels of rotation and partial occlusion of the face. He used Local Binary Patterns (LBP) as his feature extraction technique and SVMs for classification.

This research seeks to transcend and combine the two classes that FER systems are grouped into by proposing and implementing hybrid systems that first recognise AUs and subsequently use the AUs as descriptors of the six basic emotional expressions to then recognise these expressions. More importantly, it pioneers an attempt at comparing such hybrid approaches with the traditional WFE approaches of FER under the same experimental conditions. This will help determine whether taking a hybrid approach is more advantageous than the traditional approaches.

When recognising AUs, an additional question that arises is whether to use features from the entire face—global segmentation of the face—or only local regions of the face in which each specific AU is known to occur—local segmentation—during recognition. It may be that global segmentation may provide invisible but important features from the entire face that may enhance AU recognition. On the other hand, it may be that carrying out local segmentation may lead to enhanced accuracy if the recognition of AUs is truly only dependent on the region of the face within which they occur. This comparison is also carried out in this research.

The proposed research utilizes a motion-based feature extraction technique in the form of dense optical flow to characterise facial expressions and AUs, and uses SVMs to detect AUs and recognise facial expressions.



## 1.2 Research Question

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

1. Can robust autonomous hybrid FER systems be created utilising the FACS towards recognition of WFEs?
2. How do the hybrid approaches compare with traditional whole FER approaches in terms of FER accuracy?
3. How does the use of local and global segmentation of the face during feature extraction compare towards AU recognition accuracy?

## 1.3 Research Objectives

The following objectives will be met in order to answer the research questions mentioned in the previous section:

1. Implement an autonomous FER strategy that uses facial features to recognise the six basic emotional expressions.
2. Implement an autonomous FER strategy that recognises AUs and uses these with Ekman and Friesen's rules to infer and recognise the six basic emotional expressions.
3. Propose and implement hybrid FER systems that combine the two FER approaches.
4. Compare the use of global segmentation to local segmentation in terms of AU recognition accuracy.
5. Compare the hybrid and traditional approaches in terms of whole FER accuracy.

## 1.4 Premises

The following assumptions are made in this research:

- It is assumed that the user will stand or sit facing the web camera. This assumption is justified as a sign language conversation usually involves persons facing each other.
- It is assumed that only one user is present in view of the web camera at any time while performing the facial expressions. This assumption is justified as it is typical for a person to isolate themselves when conversing in loud or busy environments.
- It is assumed that the user will be of arbitrary skin colour, in front of an arbitrary background and under natural lighting. These assumptions add significant complexity to the proposed implementation but are necessary given that the SASL requires a final system that allows for the most natural setting.

## 1.5 Methodology

This research utilises the Design Science Research (DSR) methodology to help guide the modelling, implementation and analysis of components necessary for the development of the proposed FER implementations, and to help address the research questions presented in a previous section. The DSR methodology was chosen for its scientific theoretical perspective, which is necessary as this research will require a more objective approach

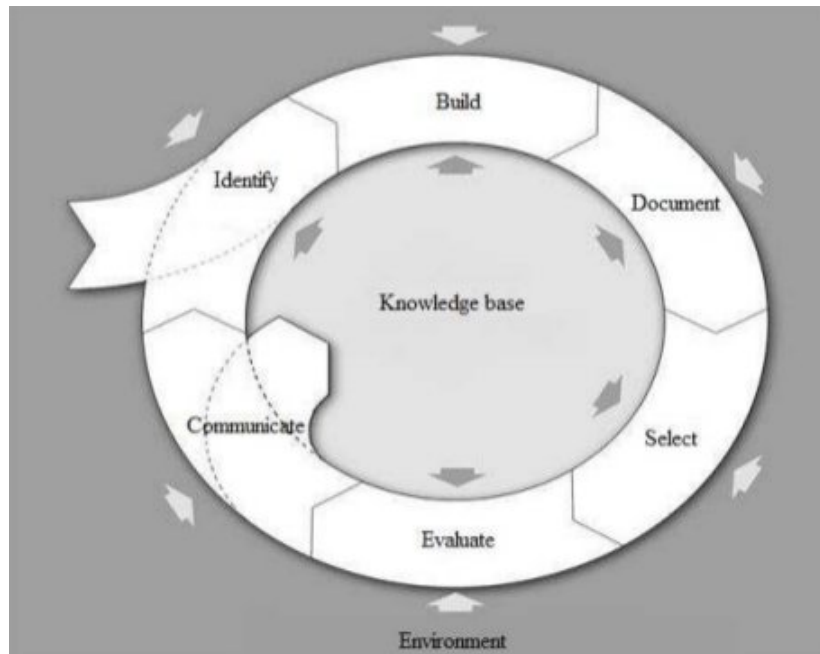


FIGURE 1.1: The DSR methodology's iterative cycle [69].

needing quantitative methods for analysis [29]. The DSR methodology's iterative cycle can be defined by its six distinctive stages depicted in Figure 1.1[69].

These stages are elaborated on and contextualised as follows:

- **Identify:** This stage refers to recognising the problem, justifying the value of a solution and defining objectives for the specific problem. In Sections 1.1 of this chapter, the problem was clearly identified and the value of a solution was motivated. Research questions that further constrained the research problem were put forward in Section 1.2 and Section 1.3 discussed the goals required to solve the research problem. Chapter 2 discusses current solutions and identifies and demonstrates the research problem in greater depth which is necessary for the identification stage. The chapter also provides a basis for the proposed implementation in subsequent stages of the methodology.
- **Build and Document:** These stages can be grouped together according to Brocke and Buddendick[69]. These stages refer to the design and development process in which an artefact capable of solving the problem—the “solution”—is developed and the representation of the artefact is created and documented. Chapter 3 discusses the methods and techniques that are used to develop the proposed artefact. Chapter 4 discusses the application of these methods and techniques in the context of this research. The chapter also represents and documents the creation of the solution in the form of clear descriptions, illustrations and flow diagrams.



- **Select and Evaluate:** These stages refer to establishing techniques to evaluate the developed solution. The evaluation criteria was described as “FER accuracy” in Section 1.2 but is further defined before testing and analysis is conducted. Chapter 5 discusses the method used to evaluate the solution and defines the evaluation criteria used in greater detail and the solution is then tested and analysed according to that method.
- **Communicate:** This stage refers to identifying the effectiveness and novelty of the solution and using the results obtained as additional requirements for a possible further iteration of the DSR cycle to solve other instances of the problem. Chapter 6 discusses the effectiveness, novelty and limitations of the proposed solution. This research limits itself to a single iteration of the DSR cycle but the chapter also puts forth recommendations for future work, thereby identifying potential areas of improvement for further iterations of the cycle in future.

## 1.6 Thesis Outline

The remainder of the thesis is arranged as follows:

**Chapter 2: *Related Work*:** This chapter discusses existing solutions in the field of FER under each of the categories of FER systems described in this chapter. This is used to further demonstrate the research problem and provide feedback into the proposed implementations.

**Chapter 3: *Image Processing Techniques for Facial Expression Recognition*:** This chapter discusses the face detection, face segmentation, feature extraction, and machine learning methods that are used in the proposed FER implementations.

**Chapter 4: *Design and Implementation of the Facial Expression Recognition Systems*:** This chapter provides an overview of the proposed FER systems and discusses the implementation of the proposed methods presented in Chapter 3.

**Chapter 5: *Experimental Results and Analysis*:** This chapter defines the techniques used to evaluate and compare the FER systems. The FER systems are then trained, tested and analysed based on the evaluation criteria in order to provide a definitive answer to the research questions.

**Chapter 6: *Conclusion*:** This chapter concludes the thesis by providing a summary of the findings from the previous chapter, highlighting the novelty, effectiveness and limitations of the proposed FER system and provides directions for future work.

## Chapter 2

# Related Work

This chapter provides an overview of existing facial expression recognition (FER) systems.

As mentioned in the previous chapter, FER systems are of three main types. The first two types are those that carry out recognition of whole facial expressions and those that recognise smaller fundamental features of facial expressions such as AUs within facial expressions. The third type of systems are those that are hybrid approaches i.e. they recognise smaller fundamental features of facial expressions and use them towards the recognition of whole facial expressions (WFEs).

The chapter will be subdivided into four sections. The first section discusses existing systems that perform AU recognition. The second section discusses existing systems that detect WFEs, referring to the six basic emotional expressions mentioned in the previous chapter. The third section discusses hybrid systems in which fundamental facial expression features are detected and then coded to recognise WFEs. Finally, the chapter is concluded by reflecting on the discussed FER systems and pinpointing the motivation that gives this research purpose.

### 2.1 Action Unit Recognition Systems

Kapoor *et al.* [37] developed an autonomous system that recognises AUs around the brow and eye region utilizing the FACS as a guideline. The system requires an Infra-Red (IR) camera to perform the preprocessing involving detecting the pupils using the red-eye effect. Once the pupils are detected, two custom templates of feature points are superimposed on the face. Figure 2.1 depicts the custom template consisting of eight

points around the contours of each eye and three points along each eyebrow, resulting in a total of 22 points describing the shape of the eyes and eyebrows.

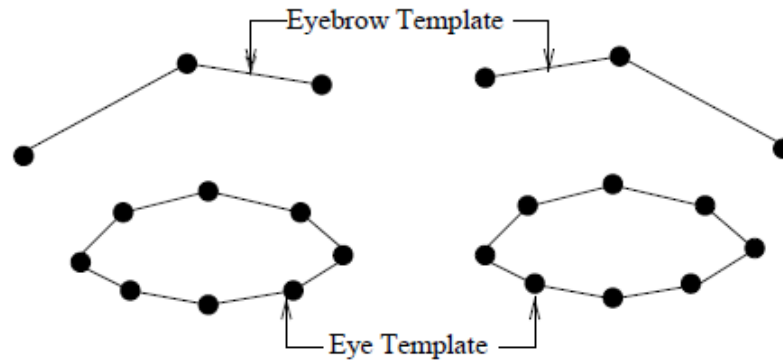


FIGURE 2.1: Custom template consisting of 22 feature points used by Kapoor *et al.* [37].

Instead of tracking the displacements of the points, the shape parameters of the eyes and eyebrows are used as feature descriptors for recognition. The AU recognition strategy uses Support Vector Machines (SVMs) as the classification method in which static frontal face images are trained and tested. The SVM is trained to recognise nine AUs or combinations of AUs along with the neutral expression. The set of AUs recognised by the system is depicted in Figure 2.2. The system also recognises head shakes and head nods by tracking the pupils through a series of frames and using the movement as input to a trained Hidden Markov Model (HMM) where five observation symbols are defined namely, Up, Down, Left, Right and None.







Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
<b>Inner Brow Raiser</b>	<b>Outer Brow Raiser</b>	<b>Brow Lowerer</b>	<b>Upper Lid Raiser</b>	<b>Cheek Raiser</b>	<b>Lid Tightener</b>

FIGURE 2.2: The set of upper face AUs recognised by Kapoor *et al.*'s system [37].

Two databases were independently sourced to evaluate the effectiveness of the system. The first database consisted of spontaneous AUs acquired by filming eight children in a real-life learning situation. The children were asked to play a game known as Fripple Place [14]. The game required the children to use mathematical reasoning while completing a variety of puzzles. Each child was given 20 minutes to work on the puzzles, and two cameras recorded the facial expressions of each child during this time. A trained FACS expert labelled the videos, indicating the AUs that were present in each frame. A total of 80 frames were manually chosen to test the system. The system achieved

an average accuracy of 61.25% in recognizing combinations of AUs and the neutral expression. Table 2.1 shows how well each combination of AUs were recognised.

Actual AUs	No. of Samples	Fully Recognised	Partially Recognised	Misses	% Full Correct
1+2	12	9	1	2	75
1+2+5	19	11	3	5	57.9
1+2+6+7	2	0	2	0	0
1+4	2	0	2	0	0
4	10	5	0	5	50
5	5	5	0	0	100
7	6	3	0	3	50
4+7	4	2	1	1	50
6+7	1	0	0	1	0
Neutral	19	14	0	5	73.7
<b>Total</b>	<b>80</b>	<b>49</b>	<b>9</b>	<b>22</b>	<b>61.25</b>

TABLE 2.1: AU recognition accuracy of Kapoor *et al.*'s system [37].

The second database consists of 10 subjects comprising an equal number of male and female subjects. The subjects were asked to shake and nod their heads while being filmed. 110 sequences were collected; 62 head nod and 48 head shake sequences. The system was then tested using the HMM for classification and received a combined average accuracy of 78.46% in recognizing head shakes and head nods. Table 2.2 shows the recognition results for head shakes and head nods using a testing sample of 65 sequences.

	Recognised	Misses
Nods	30	7
Shakes	21	7

TABLE 2.2: Head gesture accuracy of Kapoor *et al.*'s system [37].

Lien *et al.* [44] conducted a study to compare four AU recognition strategies. The strategies utilize one of three feature extraction techniques: facial feature tracking, dense flow or high gradient component detection. The strategies also utilize one of two machine learning techniques: Hidden Markov Models (HMM) or Linear Discriminant Analysis (LDA). A system overview of Lien's system is depicted in Figure 2.3.

The study aimed to detect 12 AUs following the FACS [18] guidelines. As such the face was segmented into an upper and lower region. The upper region consisted of three AUs in both the brow and eye area. The lower region consisted of six AUs in the mouth area. The descriptions and illustrations of the detected AUs are depicted in Figure 2.4. The system detected not only individual AUs but combinations of AUs as well. Before segmenting the face, normalization of all the faces in the sequence was carried out by

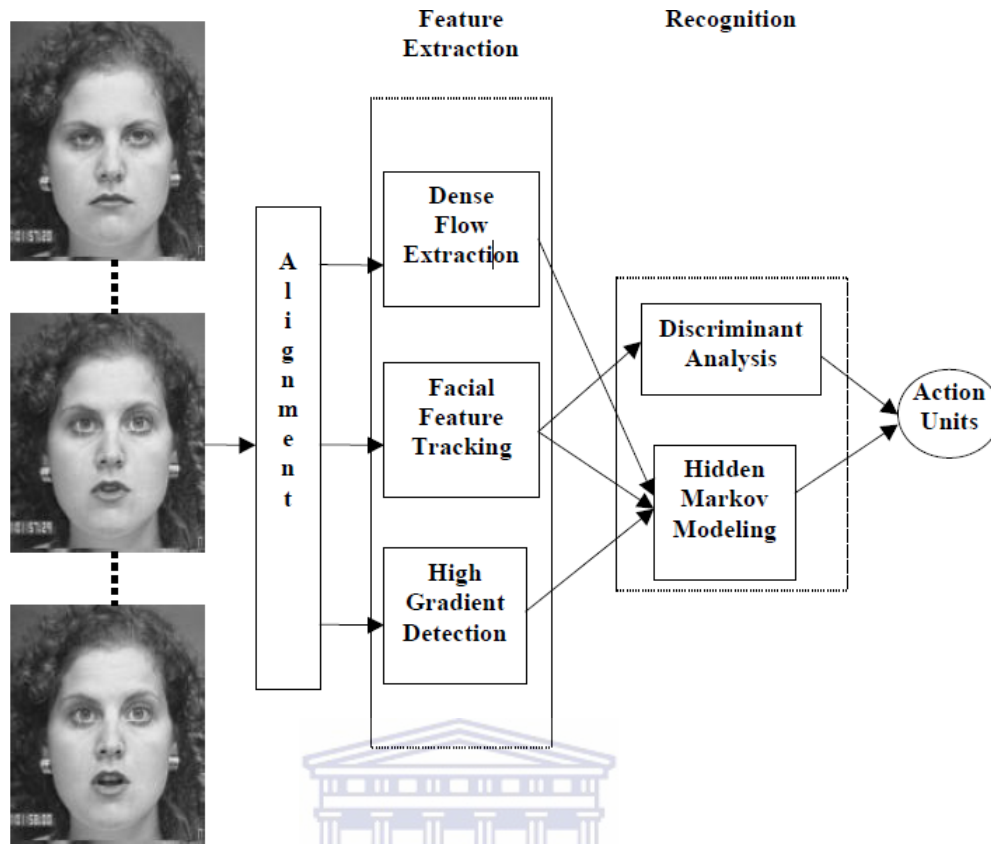


FIGURE 2.3: System overview of Lien *et al.*'s system [44].

performing a perspective transformation. This was done as expressions often occur with slight head movement. For example, a subject may raise their head when surprised. The perspective transformation looks to keep the face at the same position and orientation throughout the image sequence but this comes at a high computational cost and slows down the system quite considerably.

The Lucas-Kanade optical flow algorithm [45] was used as the facial feature tracking technique as it was described as the standard technique to estimate feature point movement efficiently. The method, however, requires that the points be manually marked in the first frame and, considering that there are 38 points that need to be marked, this manual procedure becomes very time consuming. Six points are marked around the contours of the brows, eight points around the eyes, 10 points around the mouth and 14 points around the nose. Figure 2.5 shows an example of the feature point tracking where the points are manually located on the leftmost image and thereafter automatically tracked in the consecutive frames, resulting in the activation of AU 1+2 and AU 26.

The dense flow extraction method used in the study is based on the research of Wu *et al.* [73] who developed a method to track displacement vectors using a coarse-to-fine
















Upper face Action Units					
AU4	AU1+4	AU1+2	AU5	AU6	AU7
					
Brows lowered and drawn together	Medial portion of the brows is raised and pulled together	Inner and outer portions of the brows are raised	Upper eyelids are raised	Cheeks are raised and eye opening is narrowed	Lower eyelids are raised
Lower face Action Units					
AU25	AU26	AU27	AU9+17	AU17+23+24	AU15+17
					
Lips are relaxed and parted	Lips are relaxed and parted; mandible is lowered	Mouth is stretched open and the mandible pulled down	The infraorbital triangle and center of the upper lip are pulled upwards and the chin boss is raised (AU17)	AU17 and lips are tightened, narrowed, and pressed together	Lip corners are pulled down and chin is raised
AU12	AU12+25	AU20+25			
					
Lip corners are pulled obliquely	AU12 with mouth opening	Lips are parted and pulled back laterally			

FIGURE 2.4: AUs recognised by Lien *et al.*'s system and their descriptions [44].



FIGURE 2.5: An example of the feature point tracking used by Lien *et al.* [44].

Cai-Wang wavelet representation. The wavelet model does this by representing motion vectors by a linear combination of hierarchical basis functions. The basis functions are able to alter any function into wavelet coefficients of either coarse to fine scales. The Cai-Wang dense flow is sensitive to small movements and is much more consistent when used on smoothly textured images. The biggest downfall of the Cai-Wang dense flow method, which is a very significant downfall, is its severely slow computation speed. Even when Principal Component Analysis (PCA) is used to reduce the dimensions of the flow fields, the algorithm still takes approximately 20 minutes per pair of frames when computing on an SGI-Irix workstation. An example of this dense flow extraction technique is depicted in Figure 2.6.



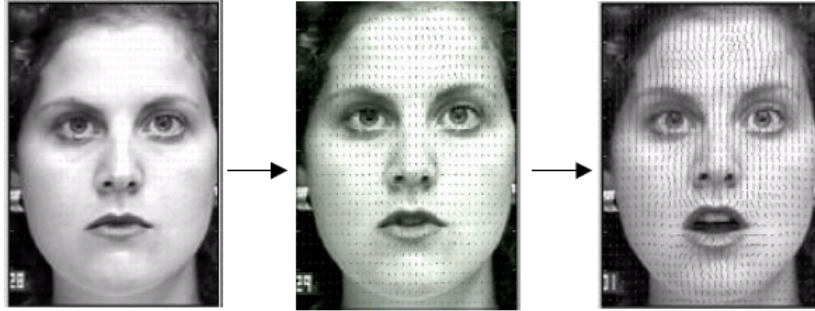


FIGURE 2.6: An example of dense optical flow computation [44].

The database used to train and test the system was independently sourced and consisted of 100 male and female adults of either European, Asian or African heritage between the ages of 18 and 35 years. The subjects sat directly in front of a camera and were asked to perform a series of facial expression sequences, each starting from the neutral expression, as is especially required for the HMM to use as its symbol sequence.

The two best performing strategies in the brow region were the dense flow tracking with HMMs which received an average recognition accuracy of 92% and facial feature point tracking with LDA receiving a recognition accuracy of 91%. The best performing strategy in the mouth region was once again the dense flow tracking with HMMs receiving an average recognition accuracy of 92%, followed by the facial feature tracking with HMMs which achieved an average recognition accuracy of 88%.

It is quite clear that dense flow with HMMs performed the best overall but required more computation, as a result of which the strategy was slower. However, it could be improved upon by substituting an affine transformation in place of the perspective transformation to align the faces. The affine transformation requires significantly less computation and, even though it is not as accurate when warping higher degrees of out of plane rotations, it performs well with relatively small movements of the head observed when performing various facial expressions. Also, the Cai-Wang dense flow algorithm can be substituted with either the Horn-Schunck [32], Lucas-Kanade [45] or Farneback [24] dense flow algorithms, all of which perform equally well in tracking movements in the face, and are more efficient at doing so.

Cohn *et al.* [11] developed a system to recognise a set of specific AUs using Ekman's FACS guidelines [18]. The system normalizes the face in the image by manually marking the medial canthus and the philtrum on the face as depicted in Figure 2.7. The system does not require any segmentation procedure given the manual normalization method implemented. The system also requires additional manual marking on the face to track 10 points in the mouth region, six in the nose region, eight in the eye region and six in the brow region resulting in a total of 30 points being manually selected on the

face for tracking. The Lucas-Kanade optical flow algorithm [45] is used to track the selected points of the face from the neutral expression to the peak of the expression. The displacement of each point is then computed by subtracting the position of the point at the peak of the expression from the initial position at the beginning of the sequence. The displacements are then separated into vertical and horizontal matrices for either the brow, eyes, nose or mouth region. The displacements are then analysed and variance-covariance matrices are created. The variance-covariance matrices are then used to predict the AUs triggered.

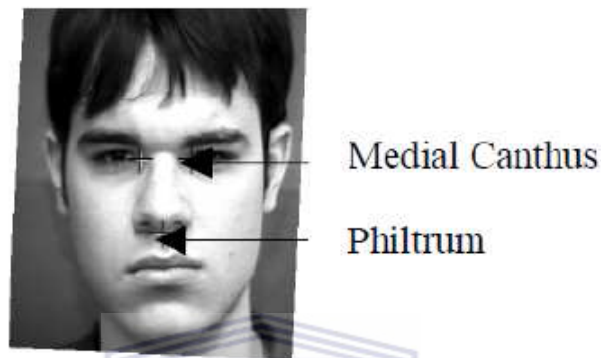


FIGURE 2.7: Features that are manually placed on the image in the normalisation method used by Cohn *et al.* [11].

The database used to train and test the system consisted of 504 image sequences. 872 AUs were confirmed to be present in the database which were acquired from 100 subjects recorded in front of a simple background. The database was randomly divided into training and cross-validation sets. The system achieved an overall recognition accuracy of 87% in detecting AUs. The system achieved an 83% accuracy in the nose and mouth region, 88% accuracy in the eye region and 92% accuracy in the brow region.

## 2.2 Whole Facial Expression Recognition Systems

Datcu and Rothkranz [13] developed a system to recognise the six basic emotions depicted in Figure 2.8 and compare the use of recognition using static images to using sequences of images for recognition. The FER system was developed to work autonomously using the Viola-Jones face detection algorithm [68] to isolate and segment the face in an image. Once the face is segmented Active Appearance Models (AAMs) are used to model each face. The AAMs carry out modelling of the face by acquiring the face shape and texture data from the image. The AAM then computes a mean face shape depicted in Figure 2.9a and mean texture illustrated in Figure 2.9b which accounts for the varied textures and shapes present in the training data. The mean face shapes and textures are



then applied to each face in the testing set so as to normalize the face, making features more accurate to track.



FIGURE 2.8: The six basic emotional expressions [13].

Both the static images and sequences of images use feature vectors containing 17 features. The 17 features pertain to the distance between modelled points on the face acquired from the AAM. The approach that uses static images – henceforth referred to as the “static approach” – only uses the frame at the peak of the expression of each sequence to extract the 17 features whereas the approach that uses sequences of images – henceforth

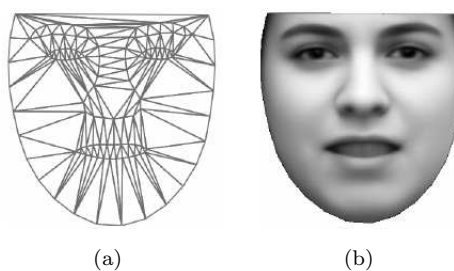


FIGURE 2.9: Active Appearance Models Used by Datcu and Rothkranz [13].

referred to as the “temporal approach” – computes the variance between each feature of the 17-dimensional feature vector, from the neutral frame to the peak of the expression.

Once the feature vectors are extracted, the system utilises an SVM to categorise the six basic emotional expressions. The Cohn-Kanade database [36] was used to train and test the system. The Cohn-Kanade database is well established and widely used in the field of AU recognition and FER. The database consists of frontal face video sequences of each of the six basic facial expressions, each progressing from the neutral expression to the peak of the expression. The number of samples used in experimentation varied for each emotion, from as little as 30 sequences for Anger to 107 sequences for Happy. A summary of the number of samples used in the experimentation can be viewed in Table 2.3.

Emotion	No. of Samples
Sadness	92
Surprise	105
Anger	30
Fear	84
Disgust	56
Happy	107

TABLE 2.3: Number of samples in the Cohn-Kanade dataset used by Datcu and Rothkranz [13].

The system achieved a recognition accuracy of 80% when trained and tested on static facial images compared to a recognition accuracy of 85% when trained and tested on sequences of facial images. Tables 2.4 and 2.5 depict confusion matrices for the static approach and temporal approach, respectively. It is quite clear that the temporal approach to emotion detection outperformed the static approach as it achieved better results for each of the six basic emotions.

Actual	Predicted(%)					
	Fear	Surprise	Sadness	Anger	Disgust	Happy
Fear	<b>84.70</b>	3.52	3.52	4.70	1.17	2.35
Surprise	12.38	<b>83.80</b>	0.95	0	0	2.85
Sadness	6.45	3.22	<b>82.79</b>	1.07	3.22	3.22
Anger	3.44	6.89	6.89	<b>75.86</b>	6.89	0
Dusgust	0	0	7.14	10.71	<b>80.35</b>	1.78
Happy	7.54	8.49	2.83	3.77	4.71	<b>72.65</b>

TABLE 2.4: Confusion matrix of the static approach of Datcu and Rothkranz [13].

One important point to be noted is that the emotion Anger was misclassified as Fear 10.71% of the time when using the temporal approach compared to 3.44% when using the

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

TABLE 2.5: Confusion matrix of the temporal approach of Datcu and Rothkranz [13].

static image approach. This suggests that there is a link between the facial movement of anger and fear over a specific time period.

Mushfieldt *et al.* [49] Developed a system that detects facial macro-expressions in the presence of rotations and partial occlusions of the face. The system is capable of detecting both frontal face images as well as side profile faces rotated to 60 degrees. As a result, a novel combined segmentation strategy that consists of two face segmentation methods, one for the frontal case and one for the rotated case, was implemented. The frontal face segmentation method employs the Viola-Jones object detection algorithm [68] to detect both the face and the eye pair in an image. Once the face and eye pair are detected, the face is segmented by taking the region formed using the height of the detected face and the width of the detected eye pair, resulting in a facial image devoid of background noise as is illustrated in Figure 2.10.

FIGURE 2.10: Isolated frontal face obtained by Mushfieldt *et al.*'s face segmentation procedure [49].

The rotated face segmentation method employs a skin detection algorithm to find the face in an image. Morphological operators such as dilate and erode are then used to distinguish the skin pixels from the pixels that might get confused with the skin. Connected Components Analysis (CCA) [5] is implemented to locate all of the skin in the image. Finally a contour map is created and the side profile of the face is isolated as can be seen in Figure 2.11.

With regard to the frontal face, a normalization technique is used to curb the effects of small misalignment variance of the head. The normalization technique detects the



FIGURE 2.11: Isolated side-view of the face obtained by Mushfieldt *et al.*'s face segmentation procedure [49].

eye positions using the eye detection algorithm of Nasiri *et al.* [52]. This highlights the eyes in the image. Once the eyes are detected, they are aligned with each other and the horizontal axis by means of an affine transformation. An illustration of the normalization procedure is depicted in Figure 2.12.

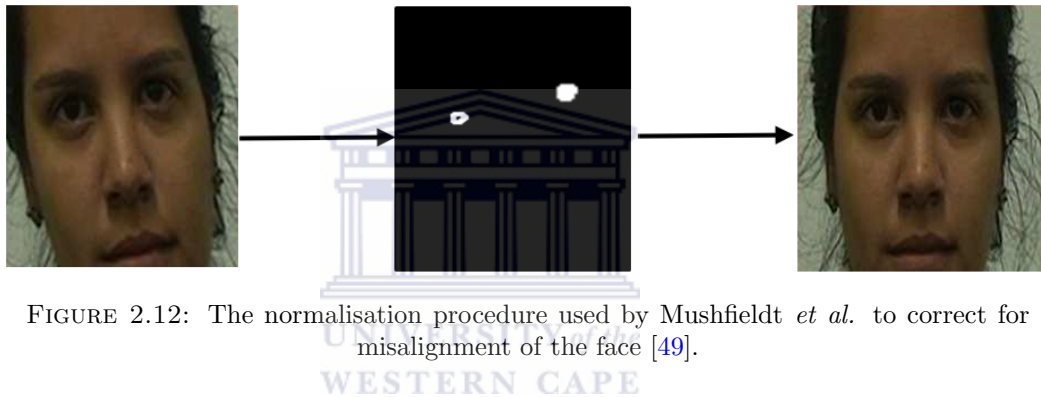


FIGURE 2.12: The normalisation procedure used by Mushfieldt *et al.* to correct for misalignment of the face [49].

Local Binary Patterns (LBPs) are used to characterise and extract the facial features of the normalized face. The features are then used as input to a multi-class SVM for training and classification of the six basic emotional expressions. The Binghamton University 3D Facial Expression (BU-3DFE) database was used to test the recognition accuracy of the system. The system achieved an average accuracy of 75% for frontal face images and achieved an average accuracy of 70% for facial images rotated 60 degrees. A summary of the average FER accuracy for each emotion using frontal and rotated faces can be viewed in Table 2.6.

Emotion	Frontal (%)	Rotated (%)
Anger	82	62
Disgust	62	52
Fear	62	50
Happy	87	95
Sadness	70	85
Surprise	90	80
<b>Average (%)</b>	<b>75</b>	<b>70</b>

TABLE 2.6: Frontal and rotated FER accuracy of Mushfieldt *et al.*'s system [49].

Schweiger *et al.* [62] developed a system to detect the six basic emotional expressions using a dense motion flow approach. The system requires the face to be manually segmented by drawing a bounding box around the face in the first frame of the video sequence. The bounding box is drawn from the top of the eyebrows to the bottom of the chin, thereby removing all sources of noise, but also requiring that the face be in the same position throughout the video sequence. It should be noted that the forehead helps convey movement in the upper half of the face so, in choosing to ignore features in the forehead, the recognition accuracy may be affected.

Once the face is manually isolated, the face is further segmented into six sub-regions by a vertical line passing through the centre of the nose and two horizontal lines, one of which passes through the centres of the eyes, and the other, across the top of the upper lip. A grid of 64 equally separated points is then superimposed on the face with each sub-region containing a subset of points. Figure 2.13 depicts an example of the segmented face with a superimposed grid of points on the face. The 64 points are then tracked through the video sequence using the Lucas-Kanade tracking algorithm [45]. The displacements of each point are then computed, followed by a calculation of the average displacement of each sub-region, resulting in a six-dimensional feature vector.



FIGURE 2.13: Superimposed grid of feature points used by Schweiger *et al.* [62].

The feature vector is then fed into a fuzzy ARTMAP Neural Network [9] to classify the six basic emotional expressions. The fuzzy ARTMAP Neural Network was chosen for its incremental supervised learning of analogue multidimensional maps. A Neural Network is created for each emotional expression whereby the average displacement of the feature vectors are evaluated and measured against the category nodes of the network. The Cohn-Kanade database was used to train and test the system. The testing phase employed a leave-one-out cross validation technique to measure the average recognition accuracy of the system.

The system ultimately achieved an average recognition accuracy of 55.84%, but this in no way reflects the true results of the system as the emotions Happy, Sadness, Surprise

and Anger achieved high recognition accuracies while Fear and Disgust received low recognition accuracies, as is illustrated by the confusion matrix in Table 2.7. Schweiger states that the results for Fear and Disgust are inconclusive, given only 8 and 10 video sequences, respectively, available to test these two expressions. It should be noted, however, that omission of the forehead could have resulted in the low accuracies of these expressions since both expressions have a significant presence in that region.

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

TABLE 2.7: Confusion matrix of Schweiger *et al.*'s classifier [62].

## 2.3 Hybrid Facial Expression Recognition Systems

Pantic *et al.* [55] developed a system to detect AUs before characterising the six basic emotional expressions using production rules. The production rules are based on the work of Ekman and Friesen [20, 21] who claim that emotional expressions can be characterised by AUs. The production rules are depicted in Table 2.8.

Emotion	AU-Coded Description (%)
Anger	4+7+(((23 or 24) with or not 17) or (16+(25 or 26)) or (10+16+(25 or 26))) with or not 2
Disgust	((10 with or not 17) or (9 with or not 17)) + (25 or 26)
Fear	(1+4) + (5+7) + 20 + (25 or 26)
Happy	6+12+16+(25 or 26)
Sadness	1+4+(6 or 7)+15+17+(25 or 26)
Surprise	(1+2)+(5 without 7)+26

TABLE 2.8: Production rules used by Pantic *et al.* to infer the six basic emotional expressions from combinations of AUs [55].

The system utilises both frontal and side view images to detect AUs. The system does not work with video sequences but instead utilizes still images. The frontal face model consists of 19 points manually inserted on the face as is depicted in Figure 2.14. Using these 19 facial points, 25 features are extracted using angles and distances between the

plotted points. The 25 features are illustrated in Table 2.9. The frontal face model consists of five additional features acquired from the shape of the mouth and chin. The mouth is represented by four specific shapes and the chin is represented by two specific shapes. A description of the five features concerned with the shape of the mouth and chin can be viewed in Table 2.10. The features are then coded to recognise a set of 27 unique AUs.

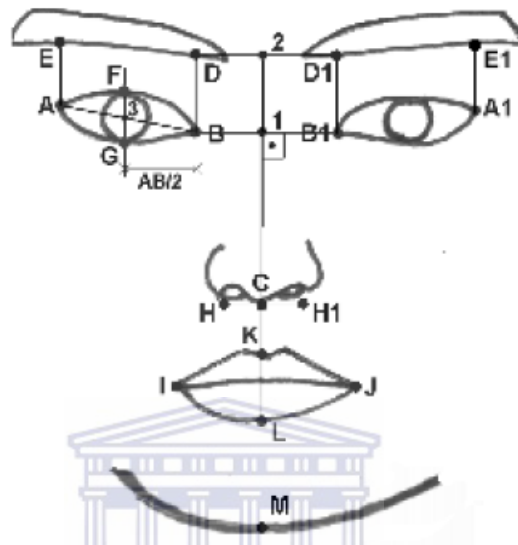


FIGURE 2.14: Facial points inserted on the frontal-view of the face as used by Pantic *et al.* [55].

The side view model consists of 10 points marked along the contours of the side profile which is depicted in Figure 2.15. The distance and curvature between the 10 points are coded to recognise 20 unique AUs. Ultimately, the models are combined to increase the quality of the face model, where the frontal view describes changes in the appearance of chin, mouth, nose, eyebrows and eyes and the side view describes changes in the appearance of the chin, jaw, mouth, nose and forehead. The combined model is also able to recognise a total of 29 unique AUs.

The system consists of two major components. The first component is known as the Integrated System for Facial Expression Recognition (ISFER) workbench. The ISFER workbench offers a wide array of feature extraction methods needed to analyse the face. The workbench first segments the face by reading in a given multi-resolution pyramid of the image and locates facial features using a raw feature map which represents a rough approximation of where the features are located. An example of the feature map is depicted in Figure 2.16. The workbench also acquires the side profile of the subject by employing Wojdel *et al.*'s [72] profile detector which uses the HSV colour space.

Once all the facial features are isolated, the workbench uses active contour models to identify the shapes of the eyes, eyebrows and mouth. The shapes are then curve



<b>Feature</b>	<b>Feature Description</b>
f1	Angle of BAD
f2	Angle of B1A1D1
f3	Distance AE
f4	Distance A1E1
f5	Distance 3F, 3 is the centre of AB
f6	Distance 4F1, 4 is the centre of A1B1
f7	Distance 3G
f8	Distance 4G1
f9	Distance FG
f10	Distance F1G1
f11	Distance CK, C is 0.5HH1 (f0)
f12	Distance IB
f13	Distance JB1
f14	Distance CI
f15	Distance CJ
f16	Distance IJ
f17	Distance KL
f18	Distance CM
f19	Image intensity in circle (r(0.5BB1), C(2)) above line (D, D1)
f20	Image intensity in circle (r(0.5BB1), C(2)) below line (D, D1)
f21	Image intensity in circle (r(0.5AB), C(A)) left from line (A, E)
f22	Image intensity in circle (r(0.5A1B1), C(A1)) right from line (A1, E1)
f23	Image intensity in the left half of the circle (r(0.5BB1), C(I))
f24	Image intensity in the right half of the circle (r(0.5BB1), C(J))
f25	Brightness distribution along the line (K, L)

TABLE 2.9: Facial features used by Pantic *et al.* to characterise frontal and rotated faces [55].

fitted to further approximate the shape of the facial features. The curve fitting utilizes mathematical techniques such as parabola functions to estimate the shape of the facial features. Figure 2.17 depicts the result of applying the active contour models. It is noted that the workbench deals only with static images. Therefore only the neutral/initial image is compared to that of the peak image.

The second major component in the FER system is referred to as the HERCULES

<b>Feature</b>	<b>Feature Description</b>
f26	Shape of lower lip when pulled downwards
f27	Mouth shape when lower lip is sucked in
f28	Mouth shape when cheeks are sucked in
f29	Circular shape of the furrows on the chin
f30	Mouth shape when the upper lip is sucked in

TABLE 2.10: Mouth and chin features used by Pantic *et al.* to characterise frontal faces only [55].



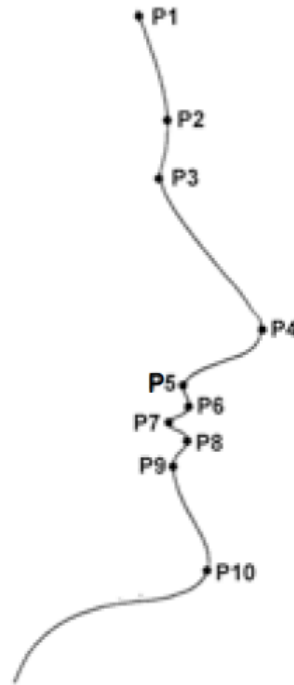


FIGURE 2.15: Facial points inserted on the side view of the face, used by Pantic *et al.* [55].

inference engine which is responsible for converting low-level face geometry into high-level AUs, followed by a conversion into the high-level weighted emotional labels. The engine utilizes a Neural Network to recognise the AUs and the six basic emotional expressions. The database used to train and test the system was independently sourced and consists of subjects who are both male and female of either European, Asian or South American heritage between the ages of 22 to 33 years. The database consists of 496 dual-view images that have been validated by eight different FACS encoders. The FACS encoders identified 31 separately activated AUs in the dataset. The system achieved an average recognition rate of 92% for the AUs present in the upper face and 86% for AUs present in the lower face.



FIGURE 2.16: An example of the feature map used by Pantic *et al.* [55].

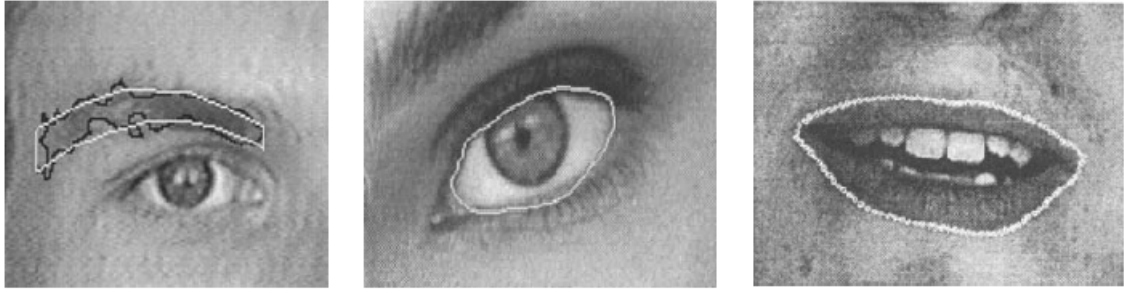


FIGURE 2.17: Active contour models computed for the eyebrows, eyes and mouth by Pantic *et al.* [55].

The emotional classification performance was tested on a set of 265 dual-view images of which 129 images contained only the six basic emotional expressions, while the remaining images contained a blend of varied emotions. The system was then trained to recognise blended emotions. The dual-views were recorded under constant illumination using a fixed light source. None of the subjects wore glasses or had a beard or moustache. The system achieved an average recognition accuracy of 91% for detecting the six basic emotional expressions and blended expressions. Table 2.11 depicts the resultant confusion matrix when the test set was run through the system.

Actual	Predicted (%)						
	Surprise	Fear	Disgust	Anger	Happiness	Sadness	Blinking
Surprise	<b>97</b>	1	0	0	0	0	2
Fear	0	<b>84</b>	0	0	0	9	7
Disgust	0	0	<b>82</b>	14	0	0	3
Anger	0	1	12	<b>84</b>	0	0	2
Happiness	1	0	0	0	<b>98</b>	0	1
Sadness	0	2	0	0	0	<b>96</b>	2
Blinking	3	1	0	0	2	1	<b>93</b>

TABLE 2.11: Confusion matrix of Pantic *et al.*'s classifier [55].

Yacoob and Davis [74] developed a FER system that employs a representation of facial feature actions. The system does not utilize the FACS as it does not detect AUs but it does detect feature actions. The feature actions are then coded to characterize the six basic emotional expressions, and an additional expression 'Blinking'. The face is first manually segmented. Six manually initialized rectangular regions are drawn on the initial frame of a sequence as is depicted in Figure 2.18. The points with the highest gradient value within the rectangular regions are then tracked throughout the sequence.

Once the points are tracked, their movements translate and scale the rectangular regions on the face through the series of frames. The rectangular regions are referred to as "windows" and the movements of these windows act as motion cues to represent basic actions such as raising, lowering or contraction of facial features. Dictionaries for

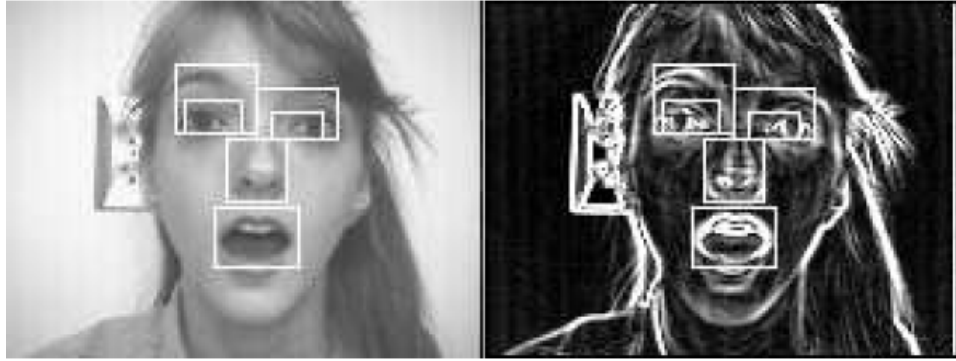


FIGURE 2.18: Manually initialised regions drawn on the face by Yacoob and Davis [74].

the brow, eyes and mouth region are created which allow the local directional motion patterns to be converted into mid-level representations for facial actions. Table 2.12 illustrates a dictionary that describes the local directional motions in the mouth region, where  $W$  denotes the rectangular window around the feature.

Component	Basic Action	Motion Cues
Upper Lip	Raising	Upward motion of $W$ 's upper part
	Lowering	Downward motion of $W$ 's upper part
	Contraction	Horizontal shrinking of $W$ 's upper part
	Expansion	Horizontal expansion of $W$ 's upper part
Lower Lip	Raising	Upward motion of $W$ 's lower part
	Lowering	Downward motion of $W$ 's lower part
	Contraction	Horizontal shrinking of $W$ 's lower part
	Expansion	Horizontal expansion of $W$ 's lower part
Left Corner	Raising	Upward motion of $W$ 's left part
	Lowering	Downward motion of $W$ 's left part
Right Corner	Raising	Upward motion of $W$ 's right part
	Lowering	Downward motion of $W$ 's right part
Mouth	Raising	Upward motion throughout $W$
	Lowering	Downward motion throughout $W$
	Compaction	Overall shrinkage in mouth's size
	Expansion	Overall expansion in mouth's size

TABLE 2.12: Yacoob and Davis' dictionary that describes the local directional motions in the mouth region, where  $W$  denotes the rectangular window around the feature [74].

The mid-level representations are then converted into emotional expressions by dividing each emotional category into three temporal components: beginning, peak and ending. Once sequences are divided into the three temporal parts, they are then modelled and used to identify other patterns that best suit the category model. The system was tested using a sample of 46 image sequences from 30 subjects, both male and females of varying skin tones. A total of 105 occurrences of the emotions recognized were present in the database. The sequences were, on average, between 8 to 16 seconds long and captured

at 30 fps. The system achieved an average recognition accuracy of 85%, receiving a recognition accuracy of 86% for Happy, 92% for Disgust, 86% for Fear, 94% for Surprise, 80% for Sadness, 92% for Anger and 65% for Blinking.

Kenji [38] developed a system similar to Yacoob and Davids in that FER is carried out using a set of predefined facial motions that are not based on the FACS. However, the system differs from Yacoob and Davids' system in that the system does not use a dictionary to convert motion cues into mid-level representations. Instead, the system uses facial muscles to identify emotions, much like the FACS.

The system first tessellates the area of the face with rectangular regions, after which optical flow feature vectors are marked within each rectangular region. A 15-dimensional feature vector is constructed and used to represent the most important points based on the variance flow throughout the video sequence. The classification of the feature vectors is then carried out using the K-Nearest Neighbours algorithm. The system was only tested on 30 sequences which were independently sourced and received an overall recognition accuracy of 80% in recognising only four emotional expressions: Happy, Disgust, Anger and Surprise.

## 2.4 Summary and Conclusion

In this chapter, the three main types of FER strategies were discussed, namely, AU recognition systems, WFE recognition systems, as well as hybrid FER systems that recognise smaller fundamental features of facial expressions such as AUs and use them towards the recognition of emotional expressions. Several conclusions can be drawn from this discussion.

The discussion demonstrated that using sequences rather than static images results in a higher recognition accuracy in most cases and is a more appropriate approach.

It was also clearly demonstrated that the dense optical flow technique is a highly accurate and efficient technique that makes an excellent feature descriptor in the proposed system.

It is also very important to note that the majority of related studies include substantial manual segmentation requirements. Input images are first manually cleared of all noise before automated processing can begin. This is time-consuming, but more importantly, limits full automation, making such approaches impractical in real-world situations.

Finally, while there are systems that used hybrid approaches, none carried out comparisons of those approaches with a version of those systems that use only the traditional approaches.

It becomes clear that it is necessary to develop a fully automatic system that can recognize WFEs based on AUs and compare it to traditional approaches under the same conditions, thus justifying this research.

The next chapter discusses the key techniques used in the motion-based approach to FER proposed in this research.



## Chapter 3

# Image Processing Techniques for Facial Expression Recognition

This chapter discusses the key techniques used in the autonomous motion-based approach to Facial Expression Recognition (FER) undertaken in this research. The proposed hybrid FER system, as well as the comparative FER systems, are completely autonomous. Hence, this chapter is subdivided into three methodological components which govern all autonomous FER approaches [39]: face detection and segmentation, feature extraction, and classification. The chapter is then concluded.

### 3.1 Face Detection and Segmentation

Face detection is the fundamental step in many autonomous FER systems as well as many learning-based gesture recognition systems [8, 51]. This is because: it can be used to pinpoint a subject in a frame; it can be used to normalize and centre a subject in the frame; and it can be used to find other objects in the frame such as the hands by using the face as a reference point. However, in this research, face detection will be used to isolate and extract the facial region from the background image.

The Viola-Jones object detection framework [68] encompasses a widely used efficient, accurate and robust implementation of face detection. As such, it has been used in the implementation of this research. The framework characterises the face using Haar-like wavelet features. The input image is converted into an intermediary image representation known as an Integral Image for faster computation of these features. The AdaBoost learning algorithm is then used to arrange a series of weak classifiers trained to recognise

various Haar-like features into a rejection cascade using a multi-tree classifier [6]. These steps are described in the following subsections.

### 3.1.1 Haar-Like Wavelet Feature Detection

Haar-like wavelet features are based on the principle of Haar wavelets and are utilized in the Viola-Jones algorithm. They are a set of two, three or four adjacent rectangular features of the same size and shape; where each rectangle is either dark or light, and are either vertically or horizontally adjacent to each other. These Haar-like wavelets are shown in the figure below (Figure 3.1).

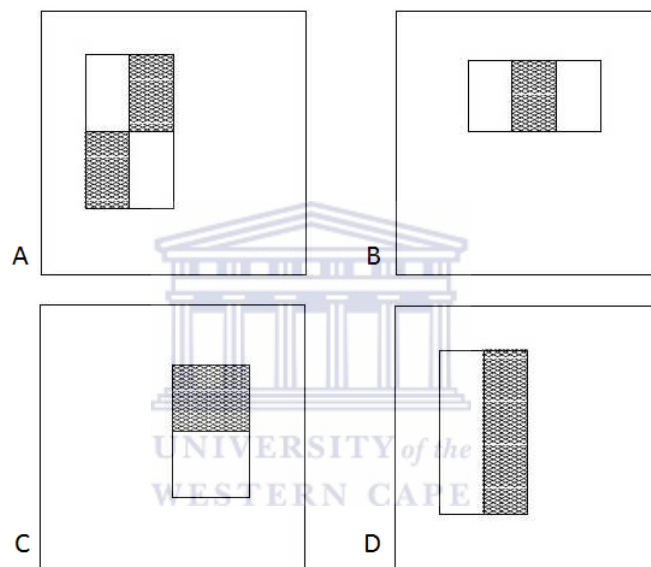


FIGURE 3.1: Haar-like features [68].

The Haar-like wavelet features are passed over an image at different scales and positions. At each scale and position, the sum of the pixels corresponding to the dark regions are subtracted from the sum of the pixels in the light regions. If the computed result surpasses an acceptable threshold value, then this specific feature is considered to be present at this scale and position.

Referring to Figure 3.1, in order to compute the four rectangular features depicted in block A, the difference amongst the sum of the pixels in the diagonal pairs of rectangles is calculated and an acceptance threshold is applied. For three rectangular features depicted in block B, the sum of the image pixels within the two outer rectangles is subtracted from the sum of the image pixels in the central rectangle and an acceptance threshold is applied. For the two rectangular features depicted in blocks C and D, the difference between the sum of the pixels within the two rectangular regions is calculated and an acceptance threshold is applied [68].

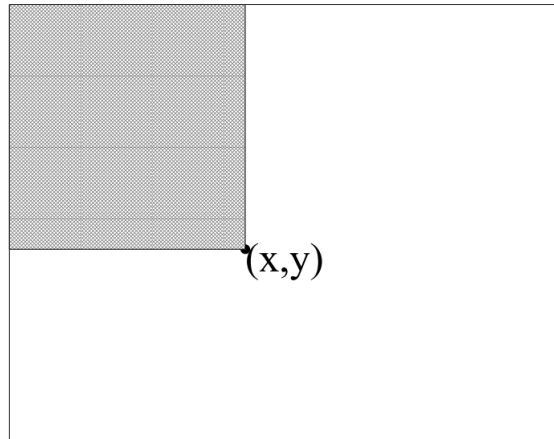


FIGURE 3.2: A visual description of the integral image representation [68].

### 3.1.2 Integral Image

Computing Haar-like wavelet features passed over an image at a variety of scales and positions can be very computationally expensive. The Viola-Jones algorithm proposes an intermediary image representation known as the Integral Image which is applied to the target image. The Integral Image representation allows for fast computation of Haar-like features at any scale and position by taking the sum of all the pixels from above and to the left of a particular pixel in a target image, as will be explained [68].

Consider a given image  $T$ , the Integral Image representation  $I(x, y)$  at any position  $(x, y)$  can be expressed in recursive form, given by the following equation:

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

It can be seen that  $I(x, y)$  is just the sum of all the pixel values to the top and left of the pixel at  $(x, y)$ , as can be seen in Figure 3.2 and further illustrated in Figure 3.3. This can also be expressed by the following equation:

$$I(x, y) = \sum_{a \leq x, b \leq y} T(a, b) \quad (3.2)$$

An example of an Integral Image computed from a target image can be seen in Figure 3.3. An Integral Image makes the computation of Haar-like features at any scale and position with only a few lookups possible. For example, in Figure 3.4, the sum of pixels in region D can be calculated by computing the sum of the Integral Image values at points 2 and 3 and subtracting it from the Integral Image value at from point 4, and adding the result to the Integral Image value at point 1. This same method is used to



compute any Haar-like feature since computing the sum of pixels in a rectangle implies the ability to compute these features.

1	2	5	1	2
2	20	50	20	5
5	50	100	50	2
2	20	50	20	1
1	5	25	1	2
5	2	25	2	5
2	1	5	2	1

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

FIGURE 3.3: An example of an integral image computed from a target image [68].

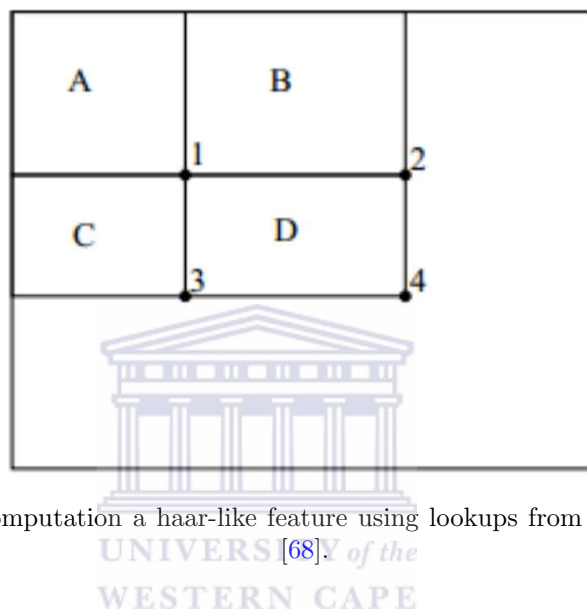


FIGURE 3.4: Computation a haar-like feature using lookups from the integral image [68].

### 3.1.3 AdaBoost Learning Algorithm

A modified AdaBoost learning algorithm was proposed by Viola and Jones [68] to improve the classification performance of a series of weak Haar-like feature classifiers. The modified algorithm boosts the recognition performance by combining the best weak classifiers to form a strong classifier in a process known as boosting. Only features that best differentiate between negative and positive examples are selected for the classifiers, thereby limiting the number of features used to create a strong classifier. Each weak classifier is assigned a particular weight where the best weighted classifier is selected after each boosting cycle.

### 3.1.4 Producing a Rejection Cascade of Weak Classifiers

The Viola-Jones algorithm constructs smaller, and hence faster, classifiers for the rejection cascade to lower the false-positive rates. Therefore, the overall performance is improved.

The rejection cascade takes the form of a degenerate decision tree, as depicted in Figure 3.5. A sub-window of the image is put through the cascade where it is tested on each weak classifier. If a negative result is obtained, it is immediately rejected which greatly reduces the computational overhead of the algorithm. Otherwise, further processing down the tree would mean a face has been correctly detected in the sub-window upto each point.

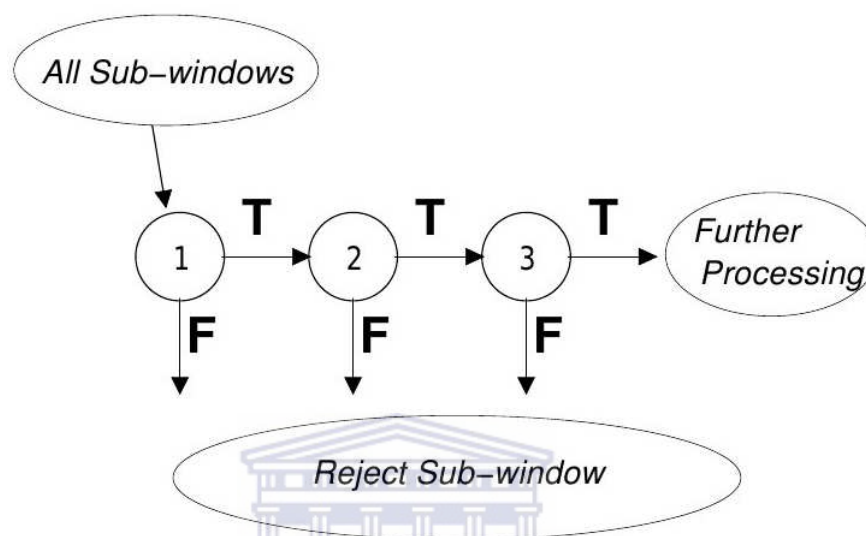


FIGURE 3.5: A rejection cascade of weak classifiers [68].

### 3.1.5 Analysis of the Viola-Jones Face Detection System

The Viola-Jones detection system's accuracy was tested on the MIT-CMU frontal face dataset. The dataset contains face images of varying sizes and subjects of varying skin tone within a variety of intricate backgrounds. Figure 3.6, depicts some examples where the Viola-Jones detection system correctly detected the faces.

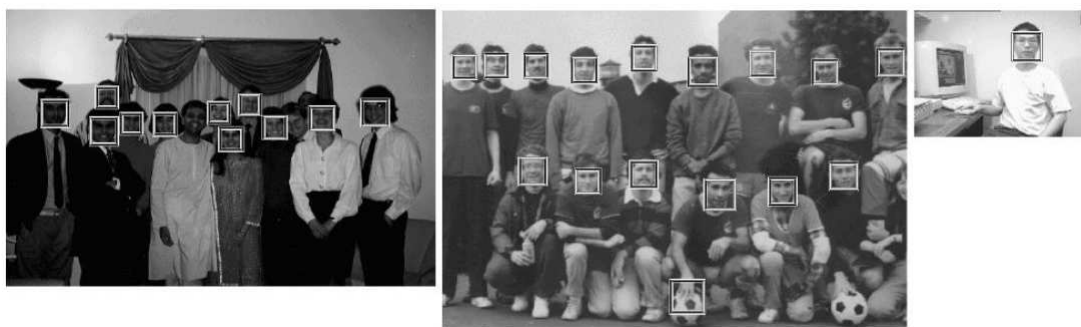


FIGURE 3.6: Various examples of the Viola-Jones face detection algorithm in operation [68].

The system was tested on 507 frontal facial images at a real-time speed of 15 fps on a 700MHz Intel Pentium 3 computer. The system received a recognition rate of 93.9%.

Due to this encouraging result, the Viola-Jones algorithm was used for the purpose of this research.

## 3.2 Feature Extraction

Dense flow was selected as the feature extraction technique for the proposed system. This motion-based feature extraction technique is ideal for recognising small muscle movements in the face.

In this section, dense flow tracking is explained, followed by a justification and explanation of the specific dense flow method used in this research.

Subsection 3.2.1 provides an overview of dense optical flow tracking. Subsection 3.2.2 describes the motion estimator used in the Farneback dense optical flow method [24] and Subsection 3.2.3 explains how the displacement of pixels are tracked by the method.

### 3.2.1 Dense Optical Flow Tracking

Dense optical flow, or dense flow for short, is a derivative of optical flow in which an estimation of the 2D displacement of an array of pixels between two adjacent frames is computed.

The aim of the procedure is to detect and track correspondence between the two frames. Dense flow therefore involves tracking correspondence at every pixel, or an array of pixels, of the frames. It also involves placing a grid of tracking points of specific density over images in which tracking should take place. The higher the number of points to track, the more features available and the higher the accuracy of motion detection, but the slower the computational speed. On the other hand, the smaller the number of points to track, the faster the processing speed, but the more sparse the resultant motion representation obtained.

The one basic assumption that is made in all optical flow methods is that the intensities (or brightness) of pixels remains constant between frames, with pixels only being translated from one position to another. This assumption can be expressed in the equation below [40]:

$$I(x, y, t) = I(x + u_x, y + u_y, t + \Delta t) \quad (3.3)$$

Where  $I(x, y, t)$  is the intensity of the pixel at  $(x, y)$  at time  $t$ ,  $(u_x, u_y)$  is the change in  $X$  and  $Y$  position of the pixel after a time period  $\Delta t$  during which the pixel moved from position  $(x, y)$  to  $(x + u_x, y + u_y)$ . The intensity of the pixel remains constant, but its position changes. This is referred to as the brightness-constancy constraint.

This constraint makes solving the problem of tracking motion in frames almost possible but it is not sufficient to infer motion completely. Additional constraints or assumptions are needed. Different optical flow methods make different assumptions. The most common assumption made is that motion across frames is smooth in the local neighbourhood of a pixel. Pixels are assumed to not “jump” or “teleport” from position to position, but move smoothly. This assumption is known as the smoothness constraint.

It is necessary to maintain a processing speed that ensures observation of the smoothness of motion in order to satisfy this constraint. For example, when tracking a single pixel, if the pixel moves by 1 pixel per frame, a processing speed of less than 1 frame per second implies that the pixel will be observed to “jump” at least 1 pixel from position to position. In order to relax this requirement to some extent, optical flow methods maintain a tracking window of specific size within which tracked pixels are assumed to reasonably move or “jump”, and beyond the bounds of which they are assumed not to move between the two frames.

The size of the tracking window depends directly on the specific target application. For applications in which very large movements are expected, the window should be large. However, attempting to locate pixels within larger tracking windows can reduce processing speeds. For applications in which small or very small motions are expected, small tracking windows are used. This results in faster processing speeds, but if pixels by chance move faster or farther than expected, tracking can be lost.

For illustrative purposes, Figure 3.7 depicts example frames of writing and running actions and their resultant motion flows. The motion flows indicate the position and direction of motion in each frame. Observing the image and motion flow in the image corresponding to the action “Writing”, the motion in the frame is correctly detected as taking place in the position where the hand is seen.

Le Bek [40] evaluated three different dense flow algorithms namely: the Horn-Schunck [32], Lucas-Kanade [45] and Farneback [24] algorithms. The algorithms were tested on two actions, namely, running and handwriting. Le Bek concluded that the Farneback algorithm produces less noisy motion flows than the Horn-Schunck and Lucas-Kanade algorithms based on results on several images.

The speed of the algorithms were compared as well and Table 3.1 shows the results of the experiment. The Lucas-Kanade method was shown to be five times faster than both

the Horn-Schunck and Farneback methods, although this is at the expense of accuracy. The speed of the Farneback method, although slower, was shown to be adequately fast, running at a real-time processing speed of 21 fps. For this reason, the Farneback dense optical flow method [24] was chosen to be used in this research.

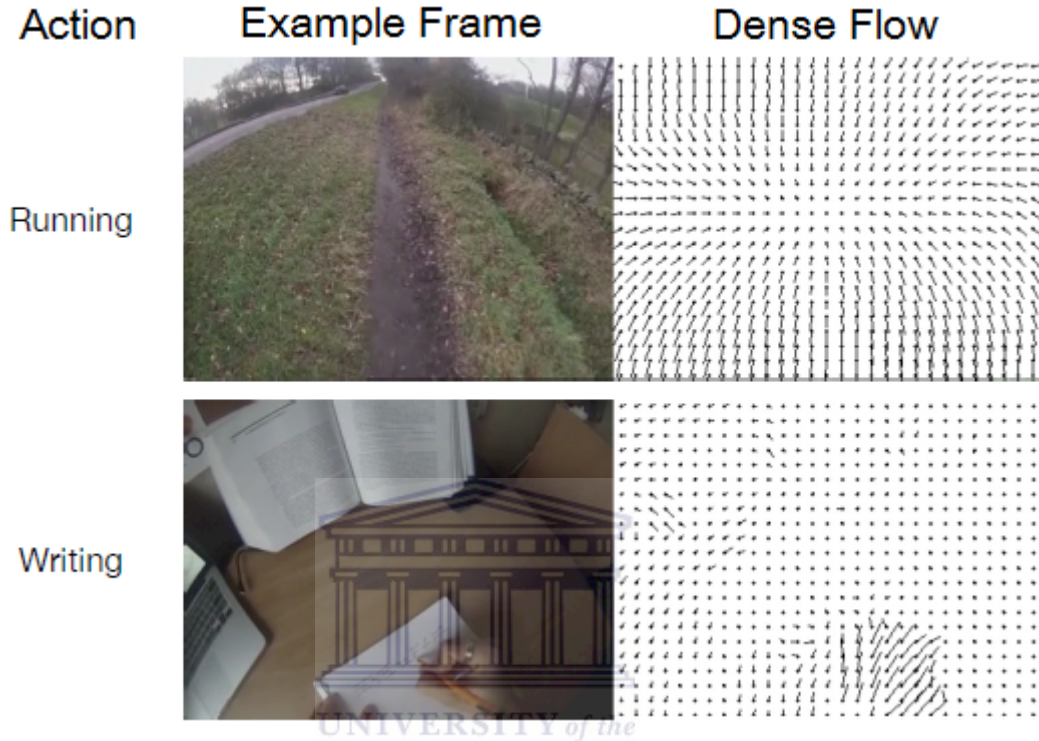


FIGURE 3.7: Example frames of running and writing and their resultant dense optical motion flows [40].

Algorithm	Running (ms)	Handwriting (ms)
Horn-Schunck	49	48
Lucas-Kanade	7	7
Farneback	53	47

TABLE 3.1: Processing time (in ms) per frame of the three dense optical flow methods compared by Le Bek, averaged over 100 frames [40].

### 3.2.2 Farneback Dense Flow Polynomial Expansion

The Farneback algorithm uses two-frame motion estimation based on polynomial expansion. The algorithm approximates the neighbourhood of each pixel in an image by a polynomial, in this case a quadratic polynomial. Thus, the local signal model in a local coordinate system is given by the following equation [24]:

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c \quad (3.4)$$

Where  $\mathbf{A}$  is a symmetric matrix,  $\mathbf{b}$  is a vector and  $c$  is a scalar. A weighted least-squares fit of the signal values in the neighbourhood is carried out in order to determine appropriate coefficient values.

### 3.2.3 Estimation of Displacement

Under this scheme, assume that a neighbourhood is approximated by the polynomial:

$$f_1(\mathbf{x}) = \mathbf{x}^T \mathbf{A}_1 \mathbf{x} + \mathbf{b}_1^T \mathbf{x} + c_1 \quad (3.5)$$

Assuming that the polynomial experiences a displacement  $\mathbf{d}$ , a new signal  $f_2$  results, defined as follows [24]:

$$f_2(\mathbf{x}) = f_1(\mathbf{x} - \mathbf{d}) \quad (3.6a)$$

$$= (\mathbf{x} - \mathbf{d})^T \mathbf{A}_1 (\mathbf{x} - \mathbf{d}) + \mathbf{b}_1^T (\mathbf{x} - \mathbf{d}) + c_1 \quad (3.6b)$$

$$= \mathbf{x}^T \mathbf{A}_1 \mathbf{x} + (\mathbf{b}_1 - 2\mathbf{A}_1 \mathbf{d})^T \mathbf{x} + \mathbf{d}^T \mathbf{A}_1 \mathbf{d} - \mathbf{b}_1^T \mathbf{d} + c_1 \quad (3.6c)$$

$$= \mathbf{x}^T \mathbf{A}_2 \mathbf{x} + \mathbf{b}_2^T \mathbf{x} + c_2 \quad (3.6d)$$

The coefficients in the quadratic equation can be equated to obtain the following:

$$\mathbf{A}_2 = \mathbf{A}_1 \quad (3.7a)$$

$$\mathbf{b}_2 = \mathbf{b}_1 - 2\mathbf{A}_1 \mathbf{d} \quad (3.7b)$$

$$c_2 = \mathbf{d}^T \mathbf{A}_1 \mathbf{d} - \mathbf{b}_1^T \mathbf{d} + c_1 \quad (3.7c)$$

It is very important to note that at this point, it becomes possible to solve the displacement  $\mathbf{d}$  as follows, if  $\mathbf{A}_1$  is assumed to be non-singular:

$$2\mathbf{A}_1 \mathbf{d} = -(\mathbf{b}_2 - \mathbf{b}_1) \quad (3.8a)$$

$$\mathbf{d} = -\frac{1}{2} \mathbf{A}_1^{-1} (\mathbf{b}_2 - \mathbf{b}_1) \quad (3.8b)$$

With the displacement acquired, it is possible to track the movements of each pixel in an image. It is also important to note that this scheme can be extended to a signal



of any dimensionality. Figure 3.8 shows the dense flow displacement computation for 2 frames.

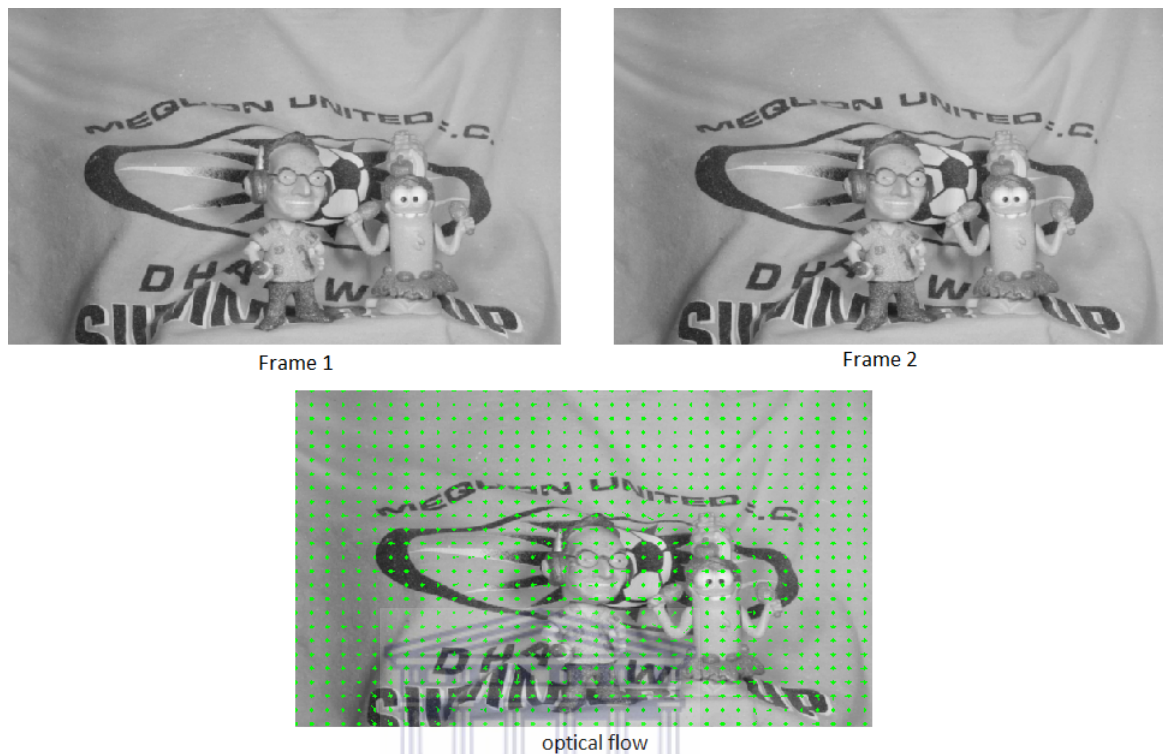


FIGURE 3.8: The dense flow displacement computation over two frames [24].

UNIVERSITY of the  
WESTERN CAPE

### 3.3 Classification

This section discusses the machine learning technique used in the proposed system. Support Vector Machines (SVMs) have been chosen for their robustness and accuracy in classifying similar patterns as will be explained.

A SVM is a supervised learning technique. SVMs are ideal for solving pattern recognition problems [8, 42, 71] and have been used extensively in sign language detection, from classifying hand shapes [42] and hand motions [3] to facial expressions [49, 71].

SVMs were first created as a binary classification technique by Vapnik [12], but were later adapted to provide for multi-class classification. SVMs have the advantage of being able to train on large images without significantly affecting training time. They have been shown to be robust, accurate and easy to train [3, 8, 12, 42, 49, 71].

Subsection 3.3.1 discusses the SVM classification process. Subsection 3.3.3 mentions types of kernels that can be used with SVMs. Finally, Subsection 3.3.4 describes various techniques utilized to solve multi-class classification problems with SVMs.

### 3.3.1 SVM Classification Process

The main purpose of a SVM is to maximize a mathematical function when given a collection of data points. It intends to find a boundary to separate data points that belong to two different classes. Figure 3.9 depicts a two-class classification problem; the red and blue points belong to two separate classes. The SVM attempts to create a boundary, known as a hyperplane, between these classes. While many different hyperplanes can be drawn to separate the two classes as shown in Figure 3.9, the SVM classification procedure chooses the hyperplane with the greatest margin between the two classes. This hyperplane is known as the “optimal hyperplane” and it is depicted in Figure 3.10.

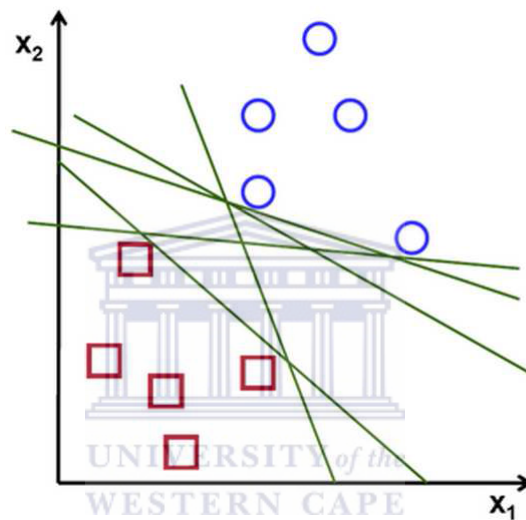


FIGURE 3.9: A two-class classification problem [12].

More formally, let the collection of  $N$  data points in Figure 3.10 be expressed as the set  $S = \{(\bar{x}_1, y_1), (\bar{x}_2, y_2), \dots, (\bar{x}_N, y_N)\}$ . Letting  $i \in \{1, 2, \dots, N\}$ , then each  $\bar{x}_i$  is a data point in  $\mathbb{R}^p$ , where  $p$  is the number of dimensions of each data point (in this example 2), and each  $y_i \in \{-1, 1\}$  is the label corresponding to each  $\bar{x}_i$ , where a value of “-1” represents the negative class, and “1” represents the positive class.

The subset  $S^+ = \{(\bar{x}_i, y_i) | y_i = 1\}$ , the positive class, corresponds to the blue points in Figure 3.10 and  $S^- = \{(\bar{x}_i, y_i) | y_i = -1\}$ , the negative class, corresponds to the red points in the figure, which are linearly separable. A hyperplane separating the two classes can be described as follows:

$$f(x) = \bar{w} \cdot \bar{x} + b = 0 \quad | \quad b \in \mathbb{R}, \bar{w} \in \mathbb{R}^p \quad (3.9)$$

Where  $\bar{w}$  is the normal vector and  $b$  is the interim term. Vector  $\bar{w}$  is a linear combination of data points  $\bar{x}_i$  and weights  $\alpha_i$ , given by:



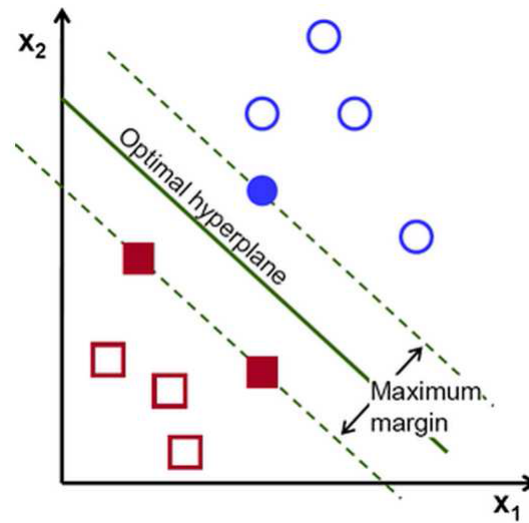


FIGURE 3.10: The optimal hyperplane that separates the two classes with a maximum margin [12].

$$\bar{w} = \sum_{i=1}^N \alpha_i \cdot \bar{x}_i \cdot y_i \quad (3.10)$$

As mentioned before, the optimal hyperplane is the boundary with the greatest margin between the two classes. Now, data points closest to the optimal hyperplane contained in either  $S^+$  or  $S^-$  are called support vectors. All support vectors satisfy either of two equations, depending on whether they are in the positive or negative subset of examples. For support vectors  $\bar{x}_+$  in the positive subset  $S^+$ , it holds that:

$$\bar{w} \cdot \bar{x}_+ + b = 1 \quad (3.11)$$

For support vectors  $\bar{x}_-$  in the negative subset  $S^-$ , it holds that:

$$\bar{w} \cdot \bar{x}_- + b = -1 \quad (3.12)$$

The distance  $d$  between support vectors from the two opposing subsets  $S^+$  and  $S^-$  is referred to as the margin and can be defined as:

$$d = \frac{2}{\|\bar{w}\|} \quad (3.13)$$

The optimal hyperplane has two important characteristics. The first property is that the data points of the two classes are clearly separated from each other [65]. Hence,

an estimation of the parameters  $\bar{w}$  and  $b$  of the optimal hyperplane have to satisfy the following:

$$y_i(\bar{w} \cdot \bar{x}_i + b) \geq 1 \quad | \quad y_i = 1 \quad (3.14a)$$

$$(3.14b)$$

$$y_i(\bar{w} \cdot \bar{x}_i + b) \leq 1 \quad | \quad y_i = -1$$

Combining the two equations 3.14 provides the following combined expression:

$$y_i(\bar{w} \cdot \bar{x}_i + b) - 1 \geq 0 \quad \forall \quad i \in \{0, \dots, N\} \quad (3.15)$$

The second characteristic of the optimal hyperplane is that the margin must be as large as possible i.e. maximum separation between points in the two classes must be achieved. Hence, the distance equation 3.13 will need to be maximized, which in turn implies a minimization of the inverse of that equation given by  $\frac{\|\bar{w}\|}{2}$ . For mathematical convenience, this implies that  $f(\bar{w}) = \frac{1}{2}(\|\bar{w}\|^2)$  can be minimized. This leads to an optimization problem in obtaining an optimal hyperplane defined as:

$$\text{Min } f(\bar{w}) \implies \text{Min } \frac{1}{2}(\|\bar{w}\|^2) \quad (3.16)$$

Which is subject to Equation 3.22. A solution to this optimization problem can be obtained by making use of the Lagrange multipliers as below:

$$L(\bar{w}, b, \alpha) = \frac{1}{2}\|\bar{w}\|^2 - \sum_{i=1}^N \alpha_i (y_i(\bar{w} \cdot \bar{x}_i + b) - 1) \quad (3.17)$$

Where  $\alpha_i$  are the Lagrange multipliers. The optimization problem is then expressed as follows:

$$\text{Max} \left[ \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^N (\bar{x}_i \bar{x}_j) \alpha_i \alpha_j y_i y_j \right] \quad (3.18)$$

which is subject to the following:

$$\sum_{i=1}^N y_i \alpha_i = 0 \quad | \quad \alpha_i \geq 0 \quad \forall \quad i \in \{0, \dots, N\} \quad (3.19)$$

The optimal hyperplane function can then be deduced as:

$$f(x) = \sum_{i \in V} \alpha_i y_i (\bar{x}_i \cdot \bar{x}) + b \quad (3.20)$$

Where  $V$  is the subset of support vectors in relation to the positive Lagrange multipliers.

### 3.3.2 Adaptation of the SVM for Non-Linear Problem Classification

The introductory theory to SVM classification in the previous subsection focused on classification problems in which the two classes in question are linearly separable. In practice, many classification problems are not linearly separable. An adaptation to the SVM that can help overcome this limitation is to transform the dataset from  $\mathbb{R}^P$  to a higher-dimensional space  $\mathcal{H}$  in which the data are linearly separable by means of a hyperplane.

Slack variables  $\xi_i$  are introduced such that the cost function becomes:

$$\text{Min}_{\bar{w}, b, \xi} \left\{ \frac{1}{2} \|\bar{w}\|^2 + C \cdot \sum_{i=1}^N \xi_i \right\} \quad (3.21)$$

which is subject to the following equation, given  $\xi \geq 0$  and  $C \geq \alpha_i \geq 0$  are constants:

$$y_i(\bar{w} \cdot \bar{x}_i + b) \geq 1 - \xi_i \quad \forall \quad i \in \{0, \dots, N\} \quad (3.22)$$

A trade-off between the margin maximisation and the classification error is controlled by the constant  $C$ .

Mercer's theorem [34] holds that the dot product of the vectors in the mapping space can be expressed equivalently as a function of the dot products of the vectors in the space. This is expressed as:

$$\begin{aligned} K(\bar{x}_i, \bar{x}_j) &= \Phi(\bar{x}_i) \cdot \Phi(\bar{x}_j) & (3.23) \\ &= (\bar{x}_i, \bar{x}_i^2) \cdot (\bar{x}_j, \bar{x}_j^2) \\ &= (\bar{x}_i \bar{x}_j) + (\bar{x}_i^2 \bar{x}_j^2) \\ &= (\bar{x}_i \bar{x}_j) + (\bar{x}_i \bar{x}_j)^2 \end{aligned}$$

where  $K(\bar{x}_i, \bar{x}_j)$  is the kernel function which can only hold true if and only if the following condition holds true for any function  $h$ :

$$\int h(\bar{x})^2 dx \text{ is finite} \implies \int K(x, y)h(x)h(y)dxdy \geq 0 \quad (3.24)$$

The selection of a kernel function that results in a transformation of the data to a higher-dimensional space in which the data are linearly separable is possible without the need to have any prior knowledge about the explicit definition of  $\Phi$ .

### 3.3.3 Kernel Functions

When data points are aggregated on a plane where two classes are non-linearly separable, kernel functions are used to map those points onto a higher-dimensional space whereby the two classes are linearly separable. Many different kernel functions based on the Mercer's theorem [34] can be used in the training and testing of the SVM, but four of the most popular are as follows:

- Radial Basis Function (RBF) kernel:  $K(\bar{x}_i, \bar{x}_j) = \exp(-\gamma\|\bar{x}_i - \bar{x}_j\|_2^2)$ , where  $\gamma > 0$
- Polynomial kernel:  $K(\bar{x}_i, \bar{x}_j) = (\gamma\bar{x}_i^T \cdot \bar{x}_j + r)^d$ , where  $\gamma > 0$
- Sigmoid kernel:  $K(\bar{x}_i, \bar{x}_j) = \tanh(\gamma\bar{x}_i^T \cdot \bar{x}_j + r)$ , where  $\gamma > 0$
- Linear kernel:  $K(\bar{x}_i, \bar{x}_j) = \bar{x}_i^T \cdot \bar{x}_j$

Where  $r, d$  and  $\gamma$  are kernel parameters. The kernel can affect the classification accuracy of an SVM and so can the parameters. Therefore, iterative optimization is needed to find the best parameter values suited to a specific classification problem. The RBF kernel has been shown to be the most accurate and widely applicable kernel function [3, 42, 49, 71]. Thus, it has been used as the kernel function in this research.

### 3.3.4 Multi-class SVM Classification Techniques

As stated previously, SVMs are fundamentally binary classifiers, limited to solving problems involving only two classes. However, various strategies have been proposed to achieve multi-class classification from a binary classification context. A comparative study done in [34] proposes three techniques to modify SVMs to solve multi-class problems. The three techniques are described in the following subsections.

### 3.3.4.1 One-Versus-All

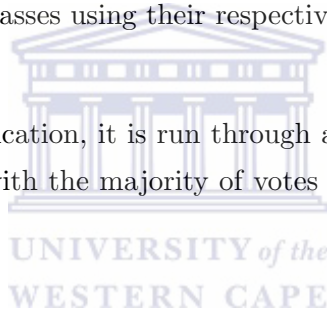
Given a  $K$ -class problem, this approach constructs  $K$  classifiers. The data points of each class  $k \in \{1, 2, \dots, K\}$  are separated from the data points of all remaining classes by an optimal hyperplane. For each case, all other classes are combined to form one class which is taken as the negative example set of class  $k$ .

If an input requires classification, it is run through all  $K$  classifiers. The predicted result is the class which obtains the maximum output value across all classifiers.

### 3.3.4.2 One-Versus-One

This technique is similar to the previous technique as it creates a series of binary classifiers. However, given a  $K$ -class problem,  $\frac{K(K-1)}{2}$  classifiers are constructed. One classifier is constructed for each pair of classes. Each classifier is trained to distinguish between the two specific classes using their respective data points as negative and positive examples.

If an input requires classification, it is run through all  $\frac{K(K-1)}{2}$  classifiers and the predicted result is the class with the majority of votes obtained using the max-wins algorithm.



### 3.3.4.3 Directed Acyclic Graph

The Directed Acyclic Graph (DAG) SVM technique was first proposed by Platt *et al.*[58]. The DAG algorithm is similar to the one-versus-one technique because, given a  $K$ -class problem,  $\frac{K(K-1)}{2}$  binary classifiers are constructed. Thereafter, a rooted binary DAG graph consisting of  $\frac{K(K-1)}{2}$  internal nodes and  $K$  leaves corresponding to each specific class is constructed.

Figure 3.11 depicts a four-class problem. Given an input that requires classification, the process starts at the root node. At this node, it is compared against classes 1 and 4. If class 4 is indicated to be the correct one then class 1 and all subsequent classifiers which contain class 1 are rejected. The input is then propagated down the remaining nodes, each time rejecting one class until only a single classifier remains. The input sequence is then said to be classified as the last remaining class. Hence, this process takes  $K-1$  steps to obtain a classification, which is significantly faster than the one-versus-one technique.

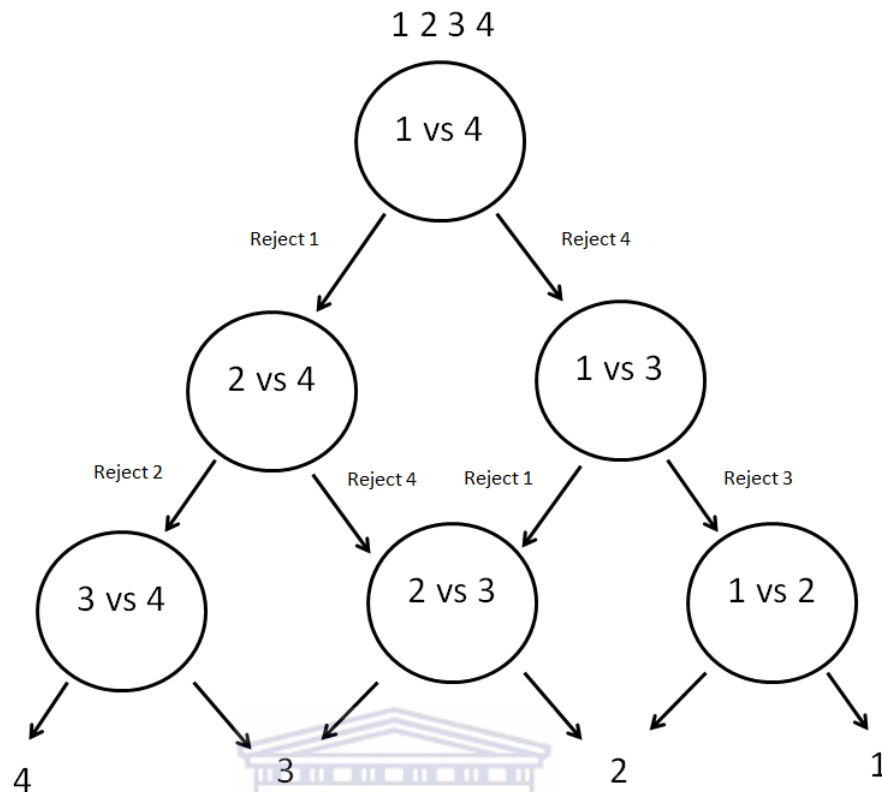


FIGURE 3.11: Directed Acyclic Graph of a 4-class problem [58].

### 3.4 Summary

In this chapter, the methodological components that form part of the autonomous FER system proposed in this research were discussed. The techniques used in the face detection and segmentation were discussed, where a detailed explanation of the Viola-Jones algorithm [68] was provided and justified.

The Farneback dense flow algorithm [24] was explained and justified as the feature extraction technique used in this research. Finally, SVMs were chosen as the machine learning technique used in this research. The classification method that SVMs employ, Kernel functions, as well as the methods devised for use with SVMs in order to solve multi-class problems were all discussed.

The next chapter discusses the use of the techniques described in this chapter towards implementing the proposed FER system.

## Chapter 4

# Design and Implementation of the Facial Expression Recognition System

This chapter discusses the design of the proposed FER systems that are compared in the next chapter and discusses the manner in which the techniques mentioned in the previous chapter were implemented. Hence, this chapter is structured in a similar manner to the previous chapter.

Section 4.1 discusses the application of the Viola-Jones algorithm to detection and segmentation of the face. Section 4.2 explains how the Farneback dense flow algorithm was used to extract the motion features from an input image sequence. Section 4.3 discusses the implementation of the four proposed FER strategies that are compared in the next chapter.

Figure 4.1 is a high-level overview of the proposed FER system. The system automatically detects and segments the face before dense flow is applied to track a neighbourhood of pixel motions. The displacement vectors are then extracted from the dense flow. Once the sequence is complete, the accumulated motion vectors are used as the input feature vector to four different FER classification methods which are explained in detail in Section 4.3.

### 4.1 Face Detection and Segmentation

This section of the system implements the Viola-Jones object detection algorithm for the purpose of isolating the face in a series of frames. The result of applying the Viola-Jones

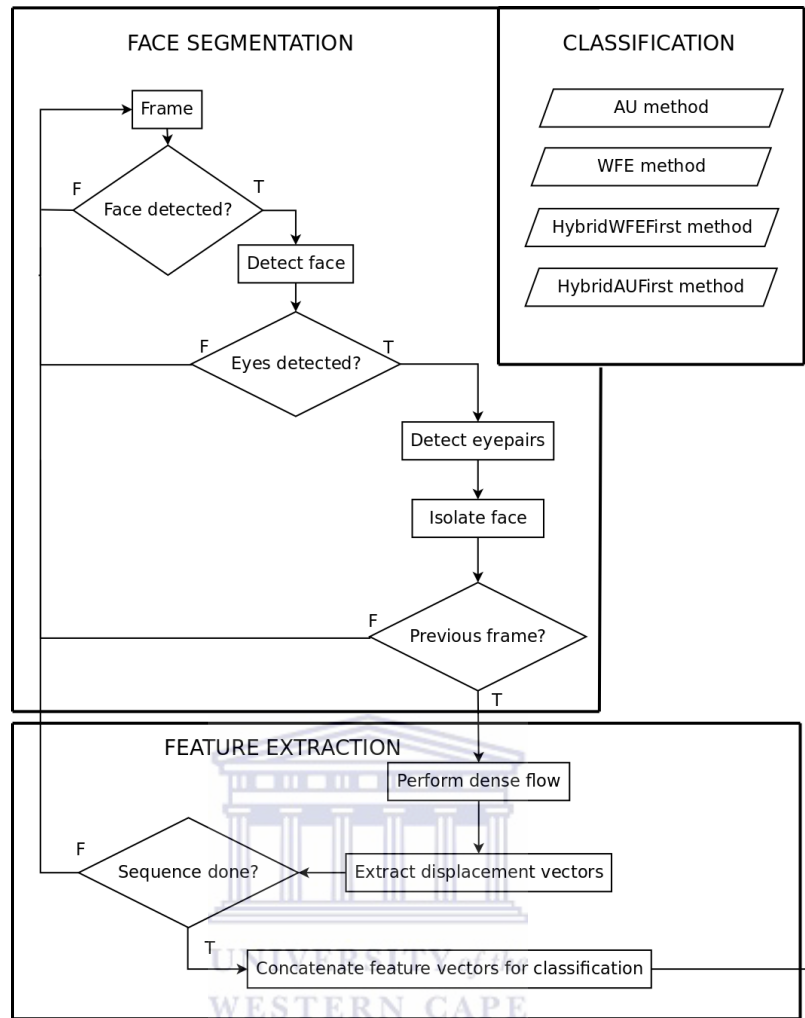


FIGURE 4.1: Processing overview of the proposed FER system.

face detection to a frame in the system is depicted in Figure 4.2.

It can be seen that the face detection does not isolate the face completely, and background noise such as ears and hair still feature in the face detected box, thereby affecting the accuracy of the system. For this reason, the face is isolated even further using the Viola-Jones algorithm with an eye-pair detection cascade to detect the eye region as depicted in Figure 4.3.

Using the facial and eye-pair boxes, the system then segments the face in three different ways to obtain regions of interest (ROIs) in which further processing and feature extraction is carried out, as required by the different FER strategies:

1. Whole face segmentation
2. Upper face segmentation
3. Lower face segmentation



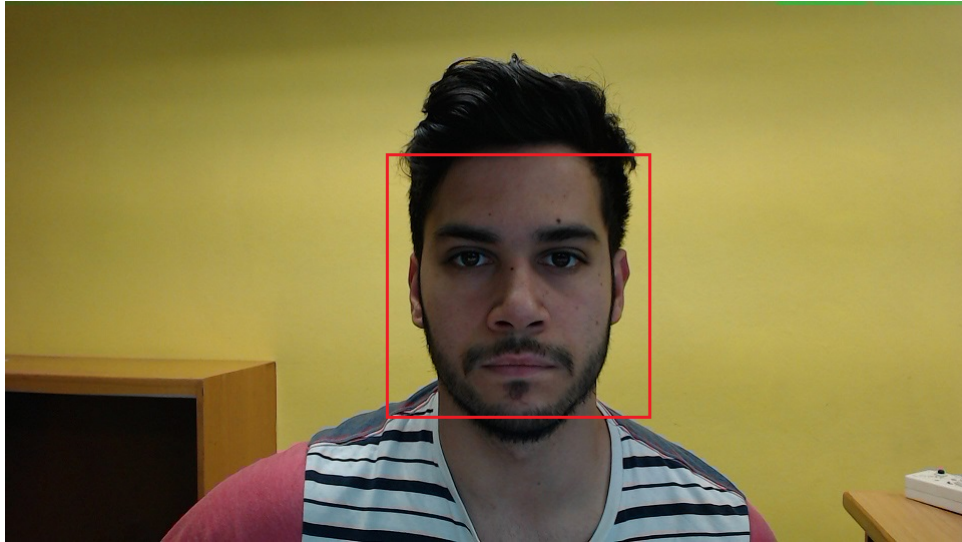


FIGURE 4.2: The Viola-Jones algorithm detects the face.

AUs occur either in the upper or lower face as stated in previous chapters. Segmenting the upper and lower face may improve the accuracy of detecting some AUs. The following subsections explain how each ROI is obtained. It should be noted at this point that once the face is detected and the segmentation process is carried out, the ROIs for the upper, lower and whole face are all obtained and resized to a resolution of  $270 \times 390$  pixels. This resolution allows for each resulting ROI to retain its compositional integrity, while allowing for faster computation of images.

It should also be noted that, henceforth, the whole face segmentation will be referred to as “global” segmentation, and both the upper and lower face segmentation will be referred to as “local” segmentation.

#### 4.1.1 Whole Face Segmentation

Once the eye region is detected, the system uses this region in conjunction with the facial box to further isolate the face. This is carried out by taking the width of the detected

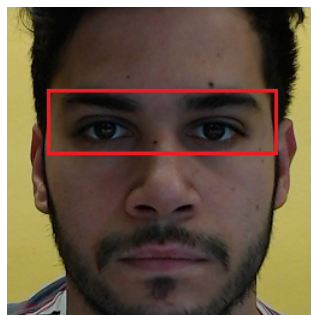


FIGURE 4.3: The Viola-Jones algorithm detects the eye-pair.

eye-pair box in Figure 4.3, along with the height of the detected facial region in Figure 4.2 as the width and height, respectively, of a segmented facial region. The resulting region is an isolated facial image with the background noise completely removed as is shown in Figure 4.4.

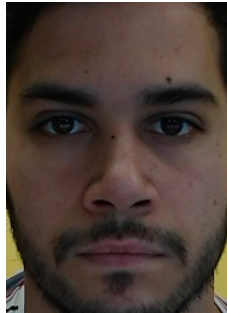


FIGURE 4.4: Isolated face after the whole face segmentation procedure.

### 4.1.2 Upper Face Segmentation

In order to segment only the upper face, the width of the detected eye-pair region in conjunction with the height consisting of the region from the top of the detected facial box to the bottom of the detected eye-pair region. The resultant segmentation is depicted in Figure 4.5. This region is used to detect AUs in the upper face.

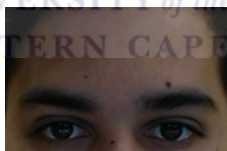


FIGURE 4.5: Isolated face after the upper face segmentation procedure.

### 4.1.3 Lower Face Segmentation

The proposed system isolates the lower face by taking the width of the detected eye-pair region coupled with the height from the bottom of the detected eye-pair region to the bottom of the detected facial box. The resultant segmentation is depicted in Figure 4.6. This region is used to detect AUs in the lower face.

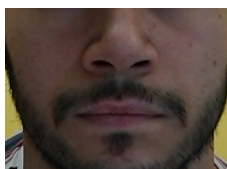


FIGURE 4.6: Isolated face after the lower face segmentation procedure.

## 4.2 Feature Extraction

The Farneback dense optical flow method computes motion vectors using pairs of sequential frames. Hence, for every frame in the sequence from the second frame onwards, the pixels of the current and previous frames are analysed and the motion vector of each point on the dense flow grid is estimated based on polynomial expansion as stated in the previous chapter. A grid implies that not all pixels in the ROI are taken into account. In the proposed implementation, a  $10 \times 10$  grid of dense flow points is placed over the ROI. This implies that every 10th row-wise and column-wise pixel is tracked and the motion vector thereof, inserted into a list of displacement vectors for further processing.

As stated previously, the ROIs are resized to a resolution of  $270 \times 390$  pixels. Hence, tracking the motion of every 10th pixel in a ROI implies that a grid of  $27 \times 39$  points is placed on each ROI as a reference descriptor of the motion in an image sequence using the Farneback dense optical flow method. This results in total of 1053 points tracked in each ROI.

The size of the tracking window used with the Farneback dense optical flow algorithm was  $15 \times 15$  pixels. The motions in the face that are to be tracked are very small. This size of tracking window was small enough to ensure real-time processing speed but was sufficiently large to ensure that pixel motions do not fall outside the tracked region.

The motion fields, being 2-dimensional vectors, have independent vertical and horizontal displacement components at each dense flow point as shown in Figure 4.7. A computation of the final feature vector of each ROI takes place as follows. For every dense flow point on the grid, a sum of all the motion vectors computed across the sequence at that point is computed. The resulting vectors are concatenated into a list and taken as the final feature vector.

Given that any input sequence to the system starts with the neutral expression and ends at the peak of a specific facial expression, the final feature vector represents the accumulated facial motion flow at each point on the dense flow grid for each specific facial expression recognised.

Mathematically, in order to compute the final feature vector  $\bar{F}$ : for each motion vector  $\bar{F}_{p,q}^n$  at each dense flow point at horizontal grid line  $p$  and vertical grid line  $q$  on the grid between frames  $n$  and  $(n - 1)$ , the sum of all the motion vectors accumulated across the entire sequence at that dense flow point  $\bar{F}_{p,q}$  is computed as follows:

$$\bar{F}_{p,q} = \sum_{n=2}^N \bar{F}_{p,q}^n \quad (4.1)$$

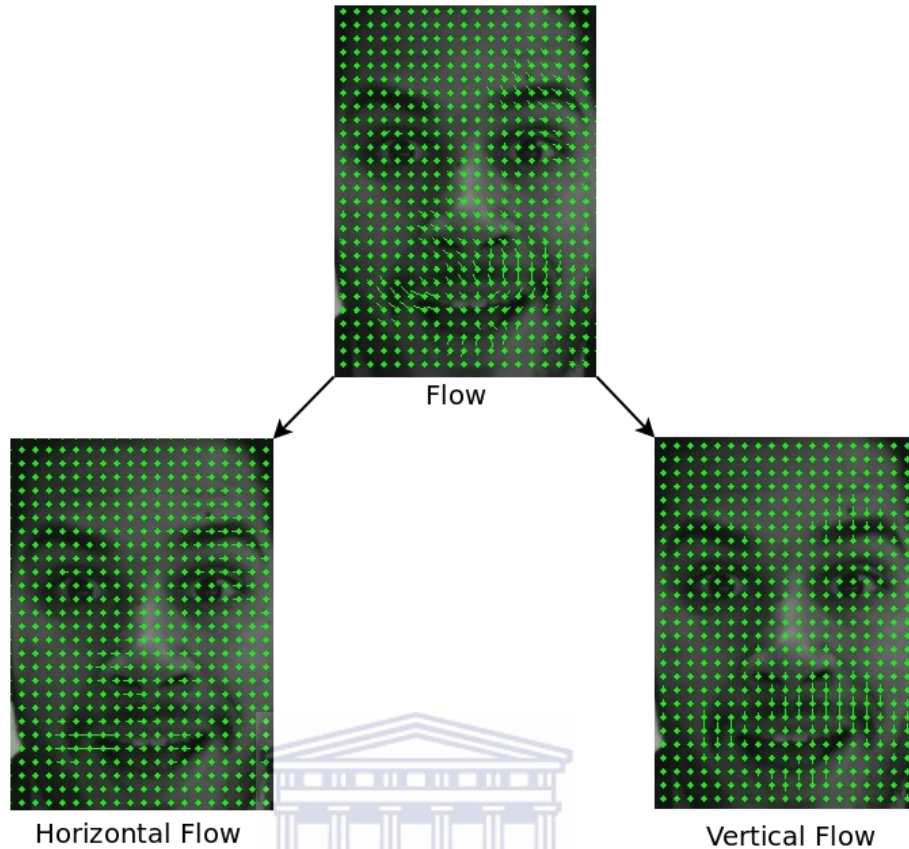


FIGURE 4.7: An example motion flow, showing its independent horizontal and vertical flows.

Where  $N$  represents the number of frames in the sequence. A concatenation of all  $\bar{F}_{p,q}$  on the dense flow grid forms the final feature vector  $\bar{F}$  as follows:

$$\bar{F} = \langle \bar{F}_{p,q} \mid p \in \{1, \dots, P\}, q \in \{1, \dots, Q\} \rangle \quad (4.2)$$

Where  $P$  and  $Q$  are the number of rows and columns, respectively, on the dense flow grid and noting that each  $\bar{F}_{p,q}$  is a vector with separate  $x$  and  $y$  motion components as follows:

$$\bar{F}_{p,q} = \langle \bar{F}_{p,q_x}, \bar{F}_{p,q_y} \rangle \quad (4.3)$$

All three ROIs produced from the aforementioned segmentation procedure have  $(P, Q) = (27, 39)$  with a total of 1053 vectors, hence, a total of 2106 features (2 features per vector). It should be noted that the feature vector produced by segmenting the face in each of the three different methods results is different, since the region of the face to be tracked is different. The procedure applied to produce these feature vectors, however, is the same. It is mentioned in later sections that the system is expected to

produce all three feature vectors for each sequence, which are used selectively by the four different FER approaches proposed and implemented. In order to differentiate between the three feature vectors, they are henceforth denoted:  $\bar{F}_W$  produced from the whole face segmentation;  $\bar{F}_U$  produced from the upper face segmentation; and  $\bar{F}_L$  produced from the lower face segmentation.

### 4.3 Classification

The first classification method—henceforth referred to as the *WFE* method for ease of reference—makes use of a multi-class SVM directly trained on the feature vector  $\bar{F}_W$  to recognise the six basic emotional expressions.

The second classification method—henceforth referred to as the *AU* method—uses a two-step procedure in which 16 relevant AUs are first recognised by means of 16 individual SVMs, each trained to recognise a specific AU, followed by the application of a set of production rules on the presence or absence of the AUs to carry out FER.

The third classification method combines the *WFE* and *AU* methods such that the prediction of the *WFE* method serves as an initial FER prediction which is then either confirmed or corrected by the *AU* method. Since the method involves first executing the *WFE* method, it is henceforth referred to as the *HybridWFEFirst* method.

The fourth classification method also combines subsets of the *WFE* and *AU* methods such that the output of the 16 individual SVM classifiers from the *AU* method are then used as features in a multi-class SVM similar to the one used in the *WFE* method to recognise the six basic emotional expressions. As such, this method is henceforth referred to as the *HybridAUFIRST* method.

The subsections that follow provide further details on each method.

#### 4.3.1 WFE Method

The *WFE* method involves training and utilizing a multi-class classifier. The classifier is trained to recognise the six basic emotional expressions, depicted in Figure 4.8. A graphical overview of the *WFE* method is depicted in Figure 4.9.

#### 4.3.2 AU Method

An overview of the *AU* method is depicted in Figure 4.10. The *AU* method involves utilizing 16 SVMs that are each trained to recognise one of the 16 AUs depicted in

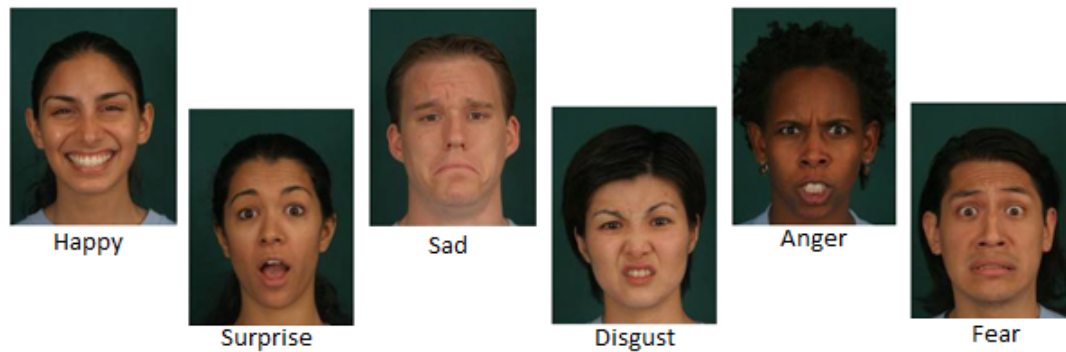


FIGURE 4.8: The six basic emotional expressions recognised.

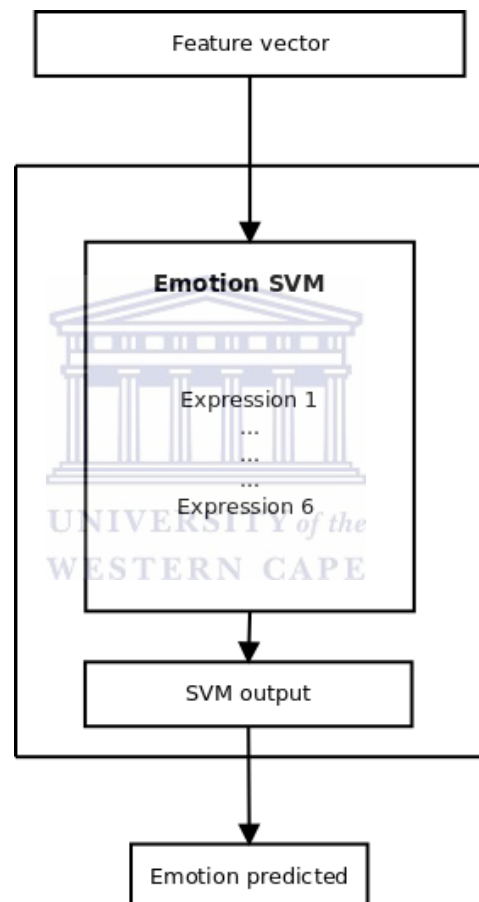
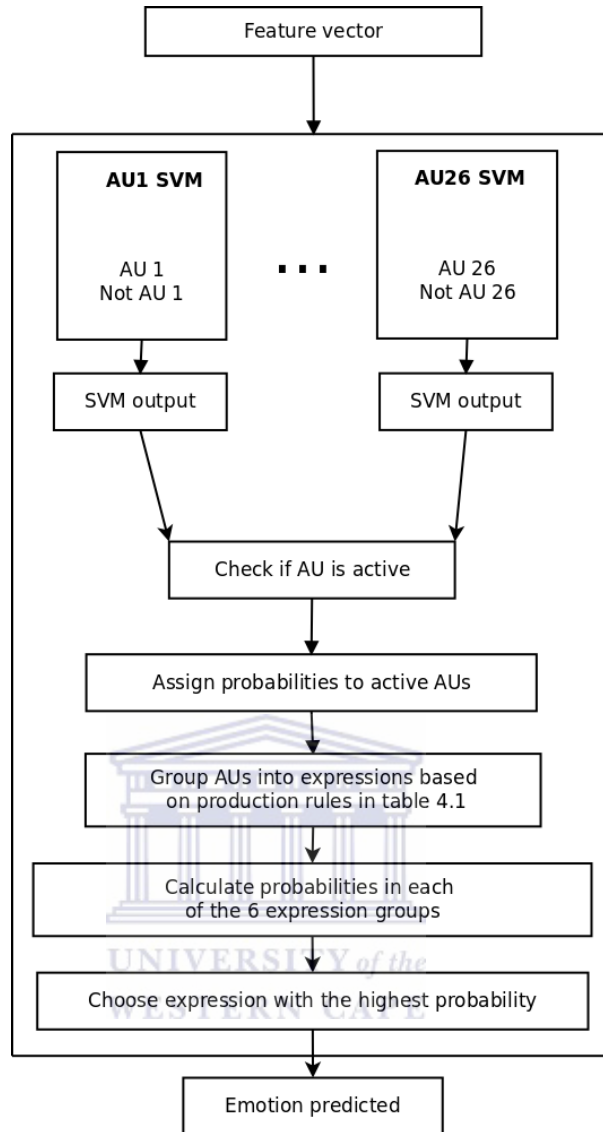
FIGURE 4.9: An overview of the *WFE* method.

Figure 4.11. The figure was provided in a previous chapter but is repeated here for ease of reference. Each AU produces a binary presence label  $l \in \{0,1\}$ , where a 0 and 1 indicate, respectively, the absence and presence of the specific AU.

Thereafter, the method uses the production rules in Table 4.1 to infer the underlying emotion in the video sequence. The table was provided in Chapter 2, but is repeated in this chapter for ease of reference. Inferring the underlying emotions is achieved by computing the sum of each production rule using the output of each relevant AU.



FIGURE 4.10: An overview of the *AU* method.

It is important to note, however, that each classifier has an intrinsic accuracy which is determined during the training procedure (explained in the next chapter) using a testing set. The accuracy of each classifier needs to be taken into account when using its prediction with the production rules.

Specifically, the prediction of a classifier of higher accuracy should have a higher weight than that of the prediction of a classifier with lower accuracy, and classifiers with the same accuracy should be equally weighted. As such, when computing the sum of each production rule, the value  $V_i$  of each classifier  $C_i$ , where  $i \in \{1, \dots, 16\}$ , is assumed to be its real-valued accuracy in the range  $[0.0, 1.0]$  determined during training, rather than its binary presence label  $l$  mentioned previously. Note that each  $V_i$  is a fixed accuracy value determined during training.

















Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
<b>Inner Brow Raiser</b>	<b>Outer Brow Raiser</b>	<b>Brow Lowerer</b>	<b>Upper Lid Raiser</b>	<b>Cheek Raiser</b>	<b>Lid Tightener</b>
Lower Face Action Units					
AU 9	AU 10	AU 12	AU 15	AU 16	AU 17
					
<b>Nose Wrinkler</b>	<b>Upper Lip Raiser</b>	<b>Lip Corner Puller</b>	<b>Lip Corner Depressor</b>	<b>Lower Lip Depressor</b>	<b>Chin Raiser</b>
AU 20	AU 23	AU 25	AU 26		
					
<b>Lip Stretcher</b>	<b>Lip Tightener</b>	<b>Lips Part</b>	<b>Jaw Drop</b>		

FIGURE 4.11: Depictions and descriptions of the 16 AUs in the upper and lower face.

Using this scheme, a fixed reference maximum value  $M_{P_j}$  is computed for each production rule  $P_j$  of each of the six expressions  $\{E_j \mid j \in \{1, \dots, 6\}\}$  by computing the sum of all  $V_i$  of the AUs referenced in each production rule  $P_j$ . This is the value indicating the highest probability of each expression  $E_j$  being present. It is implied that the maximum value  $M_{P_j}$  of each production rule may differ from that of other production rules, since it depends on the accuracies and quantities of AUs referenced in the production rule, but it is a fixed value determined once the classifiers are trained and subsequently tested using a test set. As mentioned, the procedure used to determine the accuracy of each AU classifier, and hence all  $V_i$ , is described in detail in the next chapter.

It should also be noted that although a single classifier  $C_i$  per AU  $i$  is used in the final system described here, an experiment described in detail in the next chapter was carried out to determine whether the feature vector for global or local segmentation results in a higher accuracy for each AU. For this, a comparison of the accuracy of each AU classifier given the feature vector  $\bar{F}_W$  and one of  $\bar{F}_U$  or  $\bar{F}_L$ , depending on the region of the face within which each AU occurs, was carried out.

It suffices at this point to say that in the experiment, for each AU  $i$ , two classifiers  $C_i^G$ —the classifier that uses global segmentation—and  $C_i^L$ —the classifier that uses the appropriate local segmentation—were trained and tested, and the higher performing classifier was selected as the final  $C_i$  for each AU  $i$  to be used in the *AU*, *HybridAUFIRST* and *HybridWFIRST* methods. The  $V_i$  used is that of the final  $C_i$ . At this point, it



suffices to list the  $V_i$  values of, the AU recognised by, and the type of segmentation used in, each final classifier in Table 4.2 and the  $M_{P_j}$  value of each production rule in Table 4.3.

Expression $j$	production rules $P_j$
Anger	4+7+((23 with or not 17) or (16+(25 or 26)) or (10+16+(25 or 26))) with or not 2
Disgust	((10 with or not 17) or (9 with or not 17)) + (25 or 26)
Fear	(1+4) + (5+7) + 20 + (25 or 26)
Happy	6+12+16+(25 or 26)
Sadness	1+4+(6 or 7)+15+17+(25 or 26)
Surprise	(1+2)+(5 without 7)+26

TABLE 4.1: Production rules used to infer the six basic emotional expressions using AUs [55].

Classifier $i$	AU recognised	Accuracy $V_i$	Segm. Type
1	1	0.93	Global
2	2	0.91	Local
3	4	0.80	Global
4	5	0.80	Global
5	6	0.88	Global
6	7	0.75	Local
7	9	0.87	Local
8	10	0.68	Local
9	12	0.84	Local
10	15	0.84	Global
11	16	0.72	Global
12	17	0.88	Global
13	20	0.85	Local
14	23	0.78	Local
15	25	0.88	Global
16	26	0.85	Local

TABLE 4.2: Intrinsic accuracy  $V_i$  of, and the type of segmentation (“Segm. Type”) used by, each final AU classifier.

Given an expression sequence to classify under this scheme, the sums  $S_{P_j}$  of all  $V_i$  of the AUs referenced and present in each production rule  $P_j$  are computed, which is used to compute the ratio  $R_{P_j}$  which is the ratio of the current sum  $S_{P_j}$  to the corresponding reference maximum value  $M_{P_j}$ . The expression corresponding to the highest ratio  $R_{P_j}$  is taken as the predicted expression, as this value is the predicted probability of each expression being present.

Expression $j$	Reference maximum value $M_{P_j}$
Anger	3.84
Disgust	1.76
Fear	5.03
Happy	3.33
Sadness	5.23
Surprise	3.53

TABLE 4.3: Reference maximum value  $M_{P_j}$  of the production rule  $P_j$  of each expression  $E_j$ .

### 4.3.3 HybridWFEFirst Method

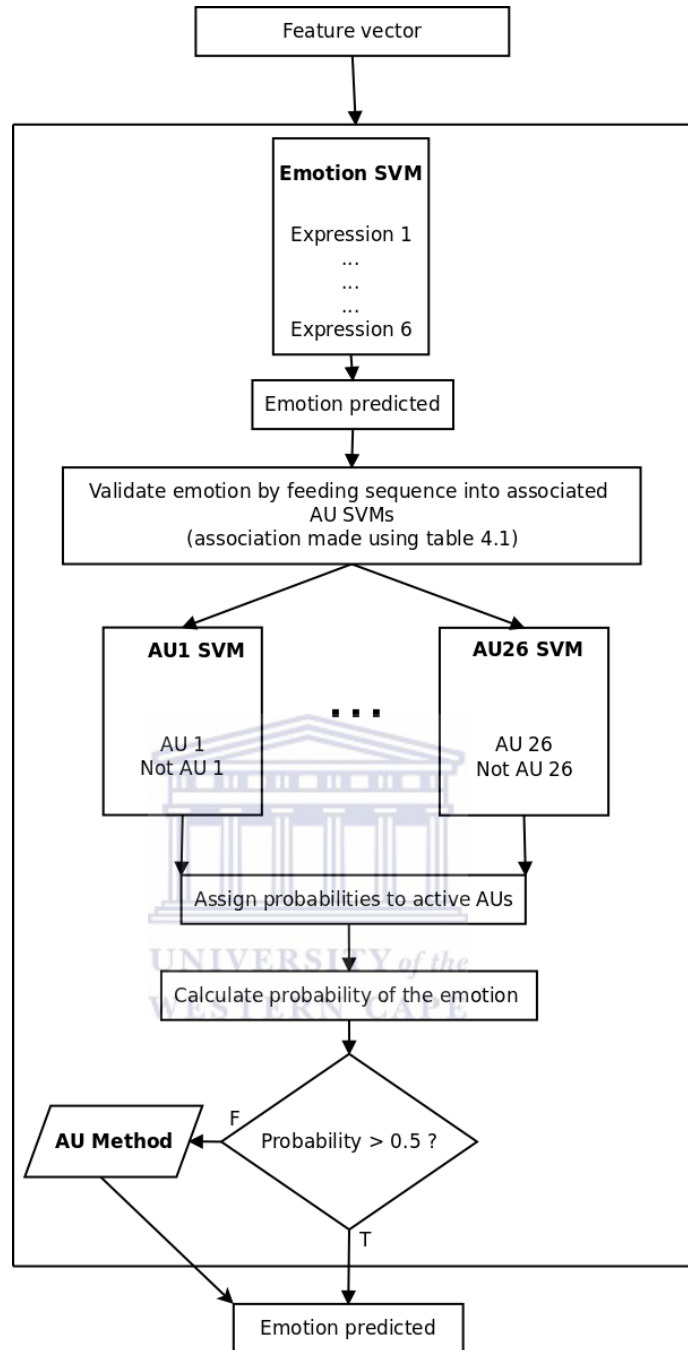
An overview of the *HybridWFEFirst* method is depicted in Figure 4.12. This method involves combining the *WFE* and *AU* methods in a unique way. A multi-class classifier is trained to recognise the six distinct emotional expressions as with the *WFE* method. The classifier receives the feature vector  $\bar{F}_W$  and uses it as input to categorize the input sequence into one of the six emotions. Once an emotion  $E_j$  is predicted, the AU classifiers associated with the predicted expression, as specified by the production rules in Table 4.1, are used to validate whether or not the relevant AUs are present or absent in order to verify whether the correct emotion has been predicted.

As per the *AU* method, a ratio  $R_{P_j}$  is computed, but initially only for the expression  $E_j$  predicted by the multi-class SVM. If the ratio value exceeds or equals a threshold of 0.5, it is concluded that the expression has been correctly detected, since at least half of the relevant AUs are determined to be present. If the value falls below the 0.5 threshold, the entire *AU* method as described before is activated whereby the feature vectors  $\bar{F}_W$ ,  $\bar{F}_U$  and  $\bar{F}_L$  are used as input to each relevant AU classifier and the ratios  $R_{P_j}$  of all expressions  $E_j$  are computed, with the expression corresponding to the highest ratio value taken as the predicted expression.

In other words, if the ratio  $R_{P_j}$  of the predicted expression  $E_j$  exceeds or equals 0.5, the method takes the form of a hybrid between the *AU* and *WFE* methods. If the ratio value falls below 0.5, the method falls back on the *AU* method only. It is therefore expected that this hybrid method should be at least as accurate as the *AU* method.

### 4.3.4 HybridAUFIRST Method

An overview of the *HybridAUFIRST* method is depicted in Figure 4.13. This method also involves training and utilizing the 16 AU classifiers to recognise the 16 AUs. This is similar to the approach taken in the *AU* method, except that the output of each classifier

FIGURE 4.12: An overview of the *HybridWFEFirst* method.

in this case is a probability  $p \in [0.0, 1.0]$  of the relevant AU occurring in the input sequence. These probabilities are then used as features in a single secondary feature vector to a multi-class SVM trained to recognise the six basic emotional expressions. Once a video sequence is fed into the system, the output probabilities  $p_i$  of all 16 AU classifiers  $C_i$ , where  $i \in \{1, \dots, 16\}$ , are computed and used to form the secondary feature vector  $\bar{F}_2$  given by:

$$\bar{F}_2 = \langle p_i \mid i \in \{1, \dots, 16\} \rangle \quad (4.4)$$

This feature vector consists of exactly 16 features representing the predicted degree of presence of each AU in the input sequence.

Given an input sequence, the final emotion prediction of this scheme is taken to be the output of the multi-class classifier similarly to the *WFE* method.

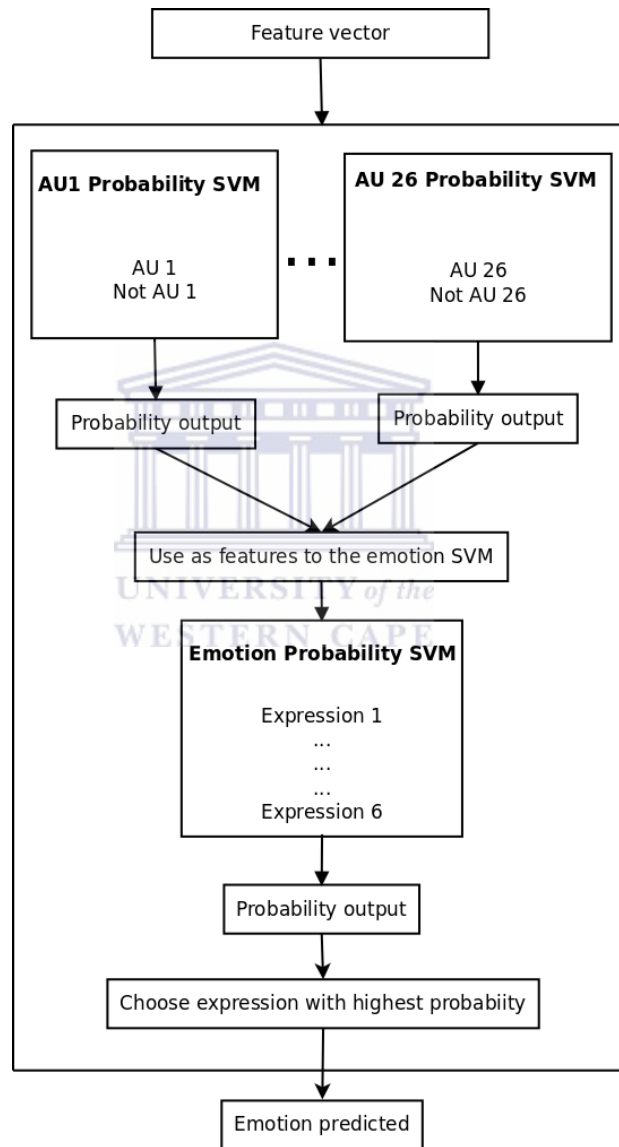


FIGURE 4.13: An overview of the *HybridAUFIRST* method.

## 4.4 Training of Classifiers

This section describes the dataset used to train the methods, the number of samples used in training each of the classifiers, the method of optimization of the classifiers, and the parameters that resulted from the optimization.

The subsections that follow are as follows: Subsection 4.4.1 describes the dataset that was selected for training and testing purposes; Subsection 4.4.2 describes the procedure used to optimize each classifier used in the system; Subsections 4.4.3 and 4.4.4 describe the training of the *AU* and *WFE* classifiers respectively.

### 4.4.1 Training Dataset

The extended Cohn-Kanade (CK+) database [46], which is an extension of the original Cohn-Kanade (CK) [36] database, was chosen as the training (and testing) dataset due to its popularity among the computer vision community and the diversity in subjects in the dataset. This database will henceforth be referred to only as “the dataset”. The database consists of 593 sequences across 123 subjects. The ages of the subjects in the dataset ranges from 18 to 50 years of age. Of the subjects, 69% are female and 31% male. In terms of ethnicity, 13% of the subjects are Afro-American, 81% are Euro-American and 6% belong to other ethnic groups.

All sequences in the dataset were recorded starting from the neutral facial expression and ending at the peak of the expression, where each peak expression is FACS coded. Only 309 sequences of the 593 sequences were labelled with an emotion. This is because these sequences are the only ones that fit the prototypic definition. Given this, only the 309 sequences could be used to recognise whole facial expressions. However, all 593 were labelled with AUs and could be used to recognise AUs.

The dataset contains frontal views as well as 30° profile views that were digitized into either  $640 \times 480$  or  $340 \times 490$  pixel arrays with 8-bit gray-scale or 24-bit Red-Green-Blue (RGB) colour values representations. Only the frontal 24-bit images were used in this research.

The next sections describe the number of sequences in which each AU is present, or associated with each emotion, in the dataset, as well as how the sequences were divided into training subsets.

#### 4.4.2 SVM Optimization Procedure

It was mentioned in Chapter 3 that the RBF kernel has been shown to be the most suitable and accurate kernel in a variety of applications. As such, it is used in this research. The RBF kernel has two parameters that can be optimized to achieve higher accuracy classification: the  $C$  and  $\gamma$  parameters. In order to optimize these parameters, a grid-search is carried out on various  $(C, \gamma)$  combinations, coupled with cross-validation [4] on a training set, in order to determine the optimum values for these parameters.

$k$ -fold cross-validation involves dividing the training set into  $k$  equal subsets. For each subset, the SVM is trained on that subset and tested on all remaining  $k - 1$  subsets to determine an accuracy. This is repeated for all  $k$  subsets, and an average accuracy—known as the cross-validation accuracy—is computed over all subsets. Given a limited training set, cross-validation provides an excellent accuracy indicator when optimizing one or more parameters.

For each  $(C, \gamma)$  combination on the grid, a cross-validation accuracy is computed. Finally, the parameter values for which the highest cross-validation accuracy is obtained is taken as the optimum combination.

This procedure was used to optimize and train all SVM classifiers across all four methods.

#### 4.4.3 Training of the AU Classifiers

Table 4.4 shows the number of sequences trained on for each AU classifier, as well as the total number of sequences available in the dataset for each AU. Only a subset of the available sequences were used for each AU training, with the remaining sequences left for testing in the next chapter. The number of sequences per AU ranged from 26–230 sequences. The large range in the number of sequences per AU is due to the fact that the database contains large variations in the number of sequences across AUs, with a large number for some AUs and a very small number for other AUs.

It is also important to note that each AU classifier was trained on an equal number of negative and positive example sequences. This was done in order to ensure an equal classification weighting between the positive (present) and negative (absent) classes for each AU classifier. Given a number of available positive training sequences for each AU, an equal number of sequences not containing the AU were selected at random from the dataset.

As previously mentioned **two** classifiers  $C_i^G$ —the classifier that uses global segmentation—and  $C_i^L$ —the classifier that uses the appropriate local segmentation—were trained for

<b>AU Recognised</b>	<b>Total Positive Examples Available</b>	<b>Positive Examples for Training</b>	<b>Negative Examples for Training</b>
1	142	86	86
2	91	55	55
4	145	87	87
5	66	40	40
6	99	60	60
7	96	58	58
9	50	30	30
10	21	13	13
12	90	54	54
15	66	40	40
16	24	15	15
17	139	84	84
20	60	36	36
23	42	26	26
25	191	115	115
26	37	23	23

TABLE 4.4: Number of positive sequences available in the dataset for each AU and the number of sequences used to train each AU classifier.

each AU  $i$  as follows: the training data was used to obtain feature vectors using both the global and local segmentation methods. For each AU, two classifiers  $C_i^G$  and  $C_i^L$  were trained. As such, a total of 32 AU classifiers were trained.

The grid-search optimization procedure mentioned in the previous subsection was applied to each AU classifier. Table 4.5 summarises the  $C$  and  $\gamma$  parameter values obtained. Once optimal parameters were determined, each classifier was re-trained on all of the training data of that AU using these parameter values in preparation for testing.

#### 4.4.4 Training of the Whole Facial Expression Multi-Class Classifiers

Table 4.6 shows the number of sequences of each emotion in the dataset, as well as the number of sequences of each emotion used to train the two multi-class classifiers of the *WFE* and *HybridAUFIRST* method. It is important to note that, for each method *WFE* and *HybridAUFIRST*, a single classifier is shared between six classes in this case. During training, it was very important to ensure balanced weighting between the six classes in each classifier by ensuring that all six classes were represented with the same number of training examples.

Unfortunately, there were only 28 sequences of the emotion “sadness” in the dataset—the lowest number of sequences out of all the emotions—at least half of which were

AU Recognised	Global Segmentation		Local Segmentation	
	$C$	$\gamma$	$C$	$\gamma$
1	512	0.00012207	512	0.00012207
2	8	0.001953125	2048	0.000305176
4	128	0.000488281	2	0.0078125
5	32	0.00012207	8	0.0078125
6	512	3.051757	128	0.000488281
7	128	0.000488281	2048	0.00122207
9	8	0.00012207	32	0.00012207
10	128	0.001953125	8	0.001953125
12	8	0.0078125	32	0.00012207
15	32	0.000488281	8	0.000305176
16	8	0.00012207	0.03125	0.001953125
17	128	0.000305176	2	0.0078125
20	2	0.001953125	8	0.001953125
23	32	0.00012207	2048	0.00012207
25	512	0.00012207	32	0.000488281
26	128	0.000488281	8	0.0078125

TABLE 4.5: Optimized parameter values for each AU classifier for global and local segmentation.

Emotion	Total Positive Examples Available	Positive Examples Used For Training
Anger	45	14
Disgust	59	14
Fear	25	14
Happy	69	14
Sadness	28	14
Surprise	83	14
<b>Total</b>	<b>309</b>	<b>84</b>

TABLE 4.6: Number of sequences available in the dataset for each emotion and the number of sequences of each emotion used to train the multi-class classifiers.

required for testing. As such, half of these examples—14 examples—were selected for training. In order to ensure balance, this dictated that all other classes were then also trained on 14 examples of each respective emotion. All remaining sequences were left for testing in the next chapter.

It should be noted that this setup, *i.e.* one classifier for six classes, is very different to that of the AU classifiers in which each classifier is trained to recognise a single AU.

Thereafter, the feature vectors used in the two classifiers were different, as explained before. For the *WFE* method, the feature vector  $\bar{F}_W$  was computed on the training data. For the *HybridAUFIRST* method, the feature vector  $\bar{F}_2$  previously described was computed on the training data.



Classifier	$C$	$\gamma$
WFE Method	32	0.00012207
HybridAUFfirst Method	128	0.00012207

TABLE 4.7: Optimized parameter values for the multi-class classifier of the *WFE* and *HybridAUFfirst* method.

Both classifiers were optimized using the grid-search optimization procedure. Table 4.7 summarises the  $C$  and  $\gamma$  parameter values obtained for each classifier. Once optimal parameters were determined, each classifier was re-trained on all of the training data using these parameter values.

## 4.5 Summary

In this chapter, a detailed discussion of the proposed FER implementations was provided. The components of the system were explained and each implementation was discussed in detail. It was explained that the face was segmented in three different ways to detect the upper face AUs, lower face AUs and whole facial expressions.

It was discussed that motion flows were extracted from an input sequence using the Farneback dense optical flow algorithm. The feature vector used was described in detail. Finally, the four different classification methods implemented in the proposed system were explained.

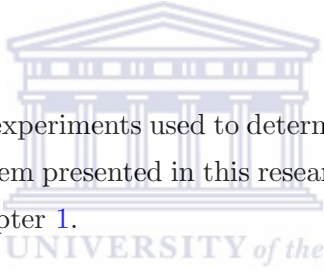
The *WFE* method utilizes a single multi-class classifier; the *AU* method makes use of 16 binary classifiers and a set of AU production rules; the *HybridWFEFirst* method uses the *WFE* method as an initial prediction which is confirmed or corrected by the *AU* method; and the *HybridAUFfirst* method makes use of a set of the 16 AU classifiers as a secondary feature generation mechanism used to train a multi-class SVM.

At this stage, it is concluded that research objectives 1, 2 and 3 set out in Chapter 1 have been successfully achieved. As a reminder, these objectives involved successfully implementing autonomous FER strategies that recognise the six basic emotional expressions using: features from the whole face; a combination of AUs and the production rules; and hybrid methodologies that were proposed.

The next chapter discusses the experiments conducted to answer the research questions.

## Chapter 5

# Design and Implementation of the Facial Expression Recognition System



This chapter discusses the experiments used to determine the recognition accuracy of the sub-components of the system presented in this research, thereby answering the research questions put forth in Chapter 1.

It describes an experiment used to determine the accuracy of the AU recognition classifiers used in three of four classification methods described in the previous chapter, including an experiment to determine whether local or global segmentation of the face is better suited to this task.

The chapter also discusses an experiment to determine the facial expression recognition accuracy of the four FER methods described in the previous chapter. A detailed analysis is performed to assess the results of each method. A comparison of the four methods is undertaken, which results in a selection of the most accurate technique for FER.

It should be noted that 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 Ubuntu 11.04 x64 operating system. Also, note that the terms “emotion” and “expression” are used interchangeably in this chapter to refer to a basic emotional expression performed in a test sequence.

The rest of the chapter is organised as follows: Section 5.1 discusses the experiment carried out to assess the accuracy of the AU recognition component of the system, as well as to determine whether global or local segmentation is more appropriate; Section

5.2 discusses the experiments carried out to assess and compare the recognition accuracy of each of the four FER methods; Section 5.3 concludes the chapter with a summary of the results.

## 5.1 AU Recognition Accuracy Experiment

This section discusses the experiment performed to assess the accuracy of the AU recognition component of the system. Three out of the four FER strategies utilize AU classifiers; thus, the accuracy of the AU recognition procedure directly affects the success of the FER strategies. Most importantly, the experiment aims to compare global and local segmentation towards AU recognition in order to obtain an answer to research question 1 set out in Chapter 1.

The following subsections describe: the criterion for a correctly recognised AU in the experiment in Subsection 5.1.1; the exact experimental procedure in Subsection 5.1.2; and a detailed discussion on the results towards determining the AU recognition accuracy and the effects of global and local segmentation on AU recognition in Subsection 5.1.3.

### 5.1.1 Criterion for a Correctly Recognised AU

Each test case involves running a test sequence through the face segmentation and feature extraction components of the system, and passing a resulting feature vector to an AU classifier. The classifier responds with a binary output that indicates whether the AU is present or absent in the input sequence. The classifier's response is compared to that of the ground truth *i.e.* whether or not that AU was actually present or absent in that sequence. If the output of the AU classifier matches the ground truth for this sequence, it is concluded as a correct classification. Otherwise, it is concluded as an incorrect classification.

### 5.1.2 Experimental Procedure

The number of total positive examples of each AU available in the dataset, as well as the number of examples used for training, were provided in the previous chapter. Table 5.1 provides the total number of positive examples of each AU used in testing. This number is, in each case, the number of positive examples in the dataset that were not used in training. As such, all the testing examples were completely unseen to the classifier. In addition, the table also specifies the region of the face—upper or lower—within which each AU occurs.

As with the training procedure, an equal number of positive and negative examples were used to test each AU. The negative examples in each case were randomly selected out of the sequences that did not contain each specific AU, and which were not used during training.

<b>AU Recognised</b>	<b>Total Positive Examples Available</b>	<b>Positive Examples for Testing</b>	<b>Negative Examples for Testing</b>	<b>Face Locality</b>
1	142	56	56	Upper
2	91	36	36	Upper
4	145	58	58	Upper
5	66	26	26	Upper
6	99	39	39	Lower
7	96	38	38	Upper
9	50	20	20	Lower
10	21	8	8	Lower
12	90	36	36	Lower
15	66	26	26	Lower
16	24	9	9	Lower
17	139	55	55	Lower
20	60	24	24	Lower
23	42	16	16	Lower
25	191	76	76	Lower
26	37	14	14	Lower

TABLE 5.1: Number of positive sequences available in the dataset for each AU and the number of sequences used to test each AU classifier.

Each of the sequences were then fed into the system and all three types of face segmentation, followed by feature extraction, were carried out on the input. Thereafter, each AU classifier was provided with both the feature vector associated with the whole face segmentation and the feature vector associated with the segmentation of the region of the face in which that specific AU occurs. Therefore, each AU classifier was provided a global and local feature vector. Thereafter, the criterion for a correctly recognised AU was applied to the output of each AU classifier.

### 5.1.3 Results and Analysis

The following subsections provide, respectively, an overview of the results and a discussion on the comparison between global and local segmentation on the AU recognition accuracy. For the rest of the discussion in this section, the term global segmentation will be abbreviated to “GS” and local segmentation to “LS” for ease of reference.

### 5.1.3.1 Overview of Results

The complete set of results obtained is provided in Table A.1 in Appendix A, an excerpt of which is provided here in Table 5.2 and depicted graphically in Figure 5.1. Note, that the AUs in the figure have been sorted in descending order of the average accuracy of the two classifiers—global and local—of each AU.

<b>AU Recognised.</b>	<b>Global Segm. Correct (%)</b>	<b>Local Segm. Correct (%)</b>
1	93	89
2	91	91
4	80	78
5	80	78
6	88	83
7	75	75
9	85	87
10	62	68
12	83	84
15	84	78
16	72	66
17	88	82
20	81	85
23	65	78
25	88	86
26	82	85

TABLE 5.2: AU classifier recognition accuracy results for the global segmentation method (“GS”) and local segmentation method (“LS”).

Referring to the figure, it is clearly seen that all 32 classifiers achieve very high AU recognition accuracies, regardless of the segmentation technique. There are no extreme outliers and all the classifiers generally perform at a consistently high accuracy level. A total of 25 of the 32 classifiers (about  $\frac{4}{5}$ th of the classifiers) achieved very high accuracies of 75% or higher, 19 classifiers (more than half of the classifiers) achieved 80% or higher, and 9 classifiers (almost a third of the classifiers) achieved 85% or higher. It is also very encouraging to note that no classifier fell below the 60% accuracy mark. This positive result indicates that, regardless of the face segmentation technique used, very high accuracy AU recognition is achieved by the proposed system.

As noted in [31], analyses of classification results of any classifier may provide an indication as to the classification decisions taken and results obtained. But, the precise cause of such results are difficult to determine, if at all possible. Therefore, an analysis to provide possible causes of these results is carried out below, noting that the most important outcome—a high recognition accuracy—has been achieved.

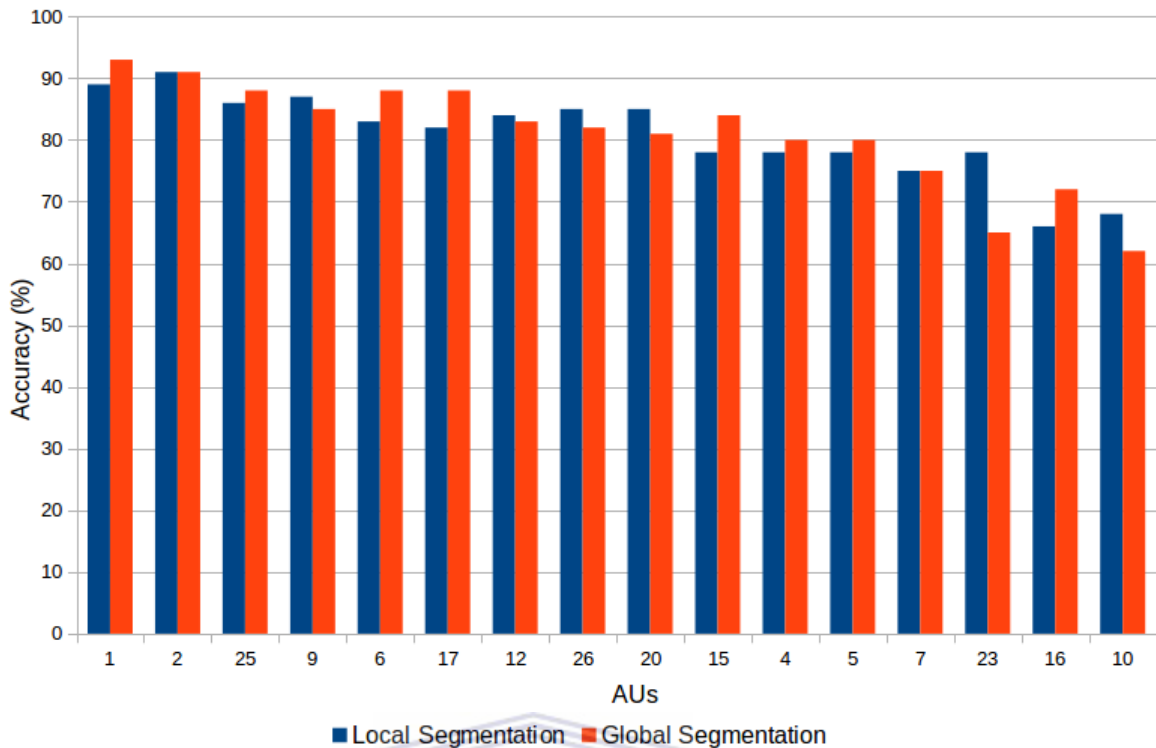


FIGURE 5.1: Graphical depiction of the AU classifier recognition accuracy results for the global segmentation method (“GS”) and local segmentation method (“LS”).

The highest accuracy obtained was for AU1, with GS, with a 93% accuracy. With LS, this classifier achieved a similarly high accuracy of 89%. AU1 pertains to the raising of the inner brows which is closely related to the emotions *Fear*, *Surprise* and *Sadness*. Figure 4.11 provided in the previous chapter is repeated here in Figure 5.2 so that the reader is able to visualise the various AUs in the context of the following discussions. The eye brows are one of the most, if not *the* most, distinct feature in the upper face, the motion of which is very easily detected. Given AU1 involves raising the inner brows, AU1 can be considered on of the least subtle<sup>1</sup> AUs in the upper face. Thus, it is expected that it would achieve the highest recognition accuracy, especially given global features.

The second highest recognition accuracy belongs to AU2 where both the GS and LS procedures received a recognition accuracy of 91%. As with AU1, AU2 involves movement of the eyebrows, in this case, the outer brows. Similar to AU1, it is expectable that AU2 would achieve such a high accuracy because of the prominence the eye brows have in the upper face region.

AU10 achieved the lowest—but by no means low—recognition accuracy of 62% for GS, although the accuracy of this AU was a higher 68% accuracy for LS. The same is true of the second and third lowest—but not low—accuracy AUs which are AU16 and AU23,

<sup>1</sup>In this and all subsequent cases, the subtlety referred to is that of motion, since it is the motion of facial features that is being tracked and used for recognition.

















Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
<b>Inner Brow Raiser</b>	<b>Outer Brow Raiser</b>	<b>Brow Lowerer</b>	<b>Upper Lid Raiser</b>	<b>Cheek Raiser</b>	<b>Lid Tightener</b>
Lower Face Action Units					
AU 9	AU 10	AU 12	AU 15	AU 16	AU 17
					
<b>Nose Wrinkler</b>	<b>Upper Lip Raiser</b>	<b>Lip Corner Puller</b>	<b>Lip Corner Depressor</b>	<b>Lower Lip Depressor</b>	<b>Chin Raiser</b>
AU 20	AU 23	AU 25	AU 26		
					
<b>Lip Stretcher</b>	<b>Lip Tightener</b>	<b>Lips Part</b>	<b>Jaw Drop</b>		

FIGURE 5.2: Depictions and descriptions of the 16 AUs in the upper and lower face.

respectively. AU23 achieves a 65% accuracy with GS which sharply increases to 78% with LS. AU16 and AU23 appear to perform better with local features. In reverse, AU16 achieves a 66% accuracy using LS which increases to 72% with GS. This AU appears to perform better with global features.

One possible reason for the lower accuracies achieved by these classifiers may be the appearance of the AUs recognised, which both involve movements of the lips. The subtlety of the muscle movements is expected to play a significant role in the recognition accuracies. AU10 refers to the raising of the upper lip found in the emotion disgust. This AU can be very subtle at times. AU16 refers to the depression of the lower lip commonly found in the emotion anger. This AU can be considered one of the subtlest movements in the lower face. It may be that the subtlety of such movements makes them less perceptible, resulting in relatively lower accuracies of the classifiers.

Another reason may be the limited number of example sequences of these AUs available for training. Figure 5.3 is a graph of the average GS and LS accuracy for each AU (on the right vertical axis) sorted in ascending order of the number of training examples available for each AU (on the left vertical axis). It can be seen that AU10 had the least number of training examples, 13 positive example sequences. While, AU16 had the second least number of training examples, 15 positive example sequences. The lowest performing classifiers were those with the least number of available positive examples in the dataset.

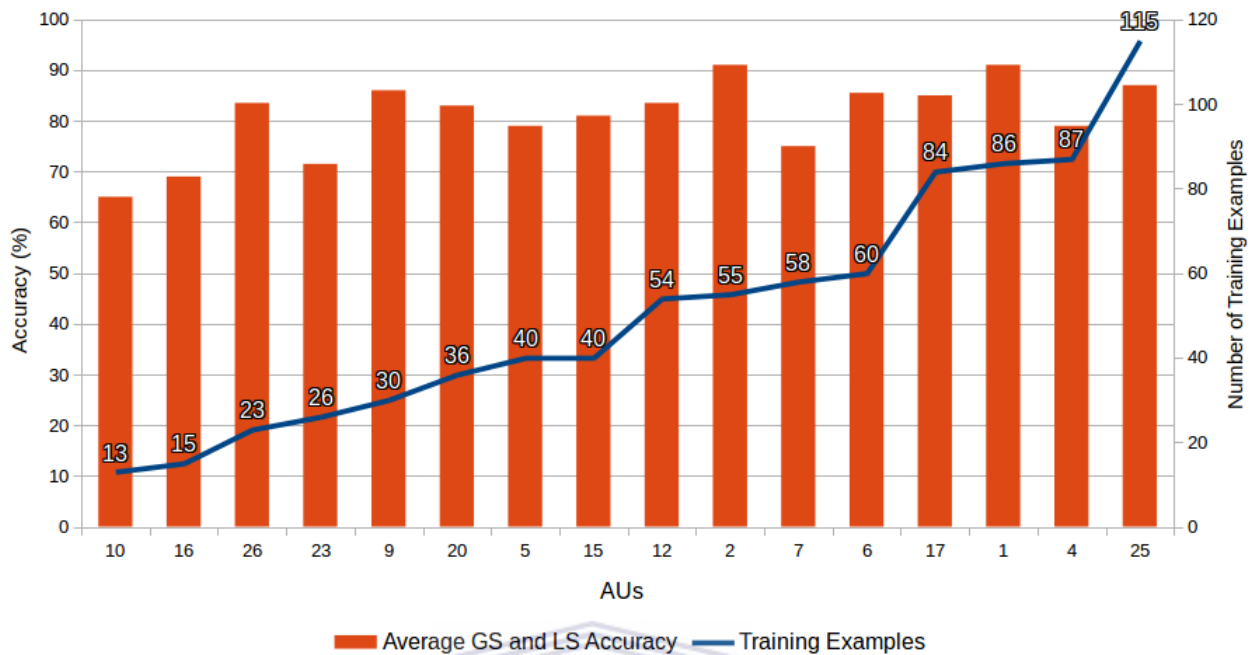


FIGURE 5.3: A graph depicting the average of the GS and LS accuracy for each AU (on the right vertical axis), sorted in ascending order of the number of training examples available for each AU (on the left vertical axis).

Given that the machine learning technique used in the classifiers of the system was the SVM, it is intuitive that some classifiers with less data would perform at a relatively lower accuracy. It is known that more training data, up to a maximum threshold, results in a more accurate SVM classifier, beyond which the accuracy stabilises. Of course, the nature of the training data is no less important, and in some cases fewer less noisy and higher quality training examples can help achieve a better accuracy [31]. In this case, AUs that involve more movement will be at an advantage as well.

In confirmation of this belief, it is further observed that the third lowest performing AU classifier—AU23—is also observed to have among the lowest (fourth lowest) number of example sequences for training, 26 positive sequences. This strengthens the belief that the number of positive sample sequences used to train the classifiers plays an important role in their recognition.

Nevertheless, as mentioned, given a small number of high quality training examples for an AU involving more movement, it is possible to achieve a very high accuracy classifier, as is the case with AU26. The fact that the AU involves a very elaborate facial movement—dropping the jaw—may put the AU at an advantage. Despite only 23



positive example sequences used for training, the classifier achieves accuracies of above 80% for both GS and LS.

For other classifiers beyond AU26, it is apparent that, for the most part, an increase in the data does not noticeably impact the classification accuracy. All of these classifiers achieve very high accuracies of above 75%. This is also in line with the explanation provided earlier.

Nevertheless, it is only possible to give indications [31]. It is most important to note that all of the classifiers achieve excellent recognition accuracies.

### 5.1.3.2 Global VS Local Segmentation for AU Recognition

Figure 5.4 is a graph of the AU recognition accuracies of each GS and LS classifier sorted in descending order of the difference between the GS accuracy and the LS accuracy of each classifier. Therefore, AU15 which has the largest difference of 6% between the GS and LS accuracy appears to the left-most extreme of the graph. AU2 and AU7 which have no difference between the GS and LS accuracy appear in the middle of the graph. Furthermore, AU23 which has the smallest difference of  $-13\%$ , *i.e.* the GS accuracy is smaller than the LS accuracy, appears to the right-most extreme of the graph.

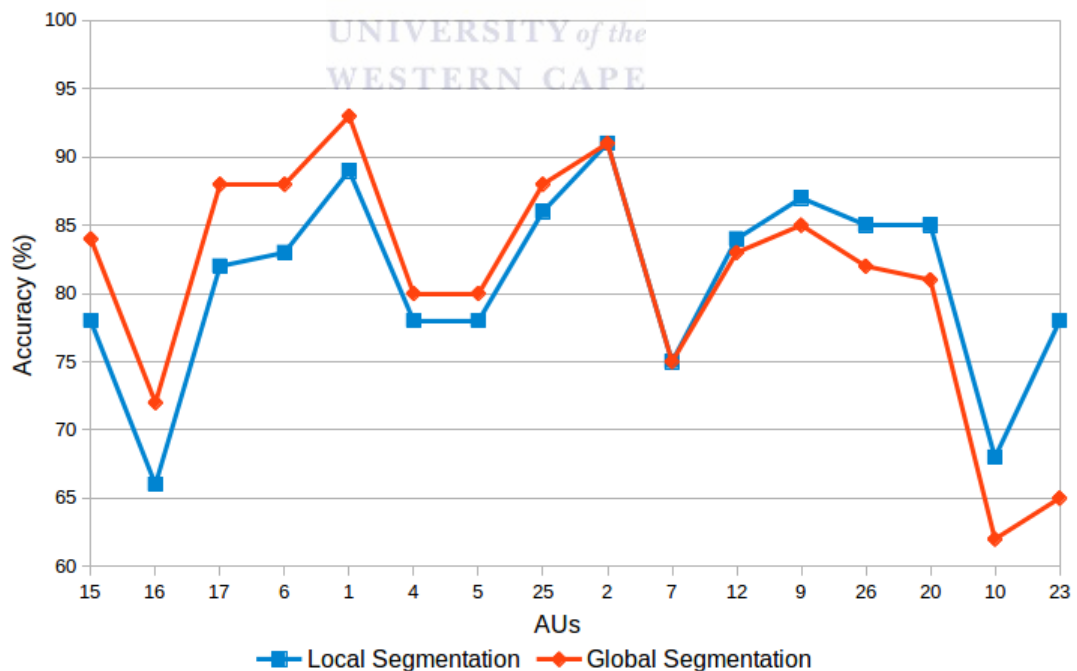


FIGURE 5.4: A graph depicting the average of the GS and LS accuracy for each AU (on the right vertical axis), sorted in ascending order of the number of training examples available for each AU (on the left vertical axis).

The graph clearly demonstrates which method—GS or LS—achieves a higher accuracy for each AU, and the extent of the difference. It is observed that the segmentation method has varied effects on various AUs.

It can be seen that the global segmentation procedure produces higher overall accuracies in recognising 8 of the 16 AUs, half of the AUs recognised. Larger differences in accuracy ranging from 4–6% are observed for 5 of these AUs, and smaller differences of 2% in the remaining 3 AUs. It is interesting to note that, of these AUs, the 3 AUs with the largest difference of 6% all involve movements in the mouth region. While it is not possible to determine exactly why these AUs benefit from GS, it is possible to say that perhaps movements in the mouth region also result in other movements in the entire face which the classifiers use to recognise these AUs.

Two AUs—AU2 and AU7—have the same accuracy for both methods. AU2 involves raising the outer brows while AU7 involves tightening the eyelids. It is expected that such movements are constrained to the upper face only. As such, using LS or GS results in the same effective features extracted and used for classification, and no observed difference between the two methods.

The remaining AUs—6 of the 16 AUs, about  $\frac{1}{3}$  of the AUs—achieve higher overall accuracies when using LS. Of these, AU23 registers a very large, in fact the largest, difference of 13% between LS and GS out of all the AU classifiers. Of the remaining AUs, 2 AUs register relatively large differences in accuracy of 4% and 6%, and 3 AUs register relatively small differences in accuracy of between 1% and 3%.

AU23 involves pursing the lips. A much higher accuracy for LS than GS may be attributed to noise in the facial images, either by default or introduced when subjects purse their lips in very different ways. Some subjects may pull their lips inwards as they purse their lips, while others may push them out, resulting in varied motions in other parts of the face. In such cases, using only the local area may result in more consistent features. The same may be true of other AUs in this group, although to a smaller extent.

It can be concluded from this discussion, in response to research question 3 which asked “How does the use of local and global segmentation of the face during feature extraction compare towards AU recognition accuracy?”, that both feature extraction techniques result in very high accuracy in AU recognition, but either technique may be somewhat more suited to some AUs than others, while there is no difference for some AUs. The choice is AU-specific.

## 5.2 Facial Expression Recognition Accuracy Experiment

This section discusses the experimentation performed to assess the accuracy of the four FER strategies and compare them; thus, achieving the final research objective 5 set out in Chapter 1 and providing an answer to research question 2 posed in Chapter 1 that is tied to this objective.

Subsection 5.2.1 describes the criterion for a correctly recognised facial expression; subsection 5.2.2 describes the exact experimental procedure carried out; and subsection 5.2.3 discusses and analyses the results obtained.

### 5.2.1 Criterion for a Correctly Recognised Facial Expression

Each test case in this experiment involves running a test sequence through the face segmentation and feature extraction components of the system, and passing the relevant feature vectors that result into each of the four FER methods. Each classifier responds with a label  $l \in \{1, \dots, 6\}$ , where each label corresponds to one of the six basic emotional expressions as detailed in Table 5.3.

Each of the sequences used in testing were also labelled in the same way. Given the output of a method, it is compared to that of the input sequence. If the two labels match, it is deemed a correct classification. If the labels do not match, this is deemed an incorrect classification.

### 5.2.2 Experimental Procedure

The number of positive examples available in the dataset for each emotion and the number of sequences of each emotion used in this experiment are provided in Table 5.3. A total of 225 sequences were used in the experimentation. As with the AU accuracy experiment, all of the examples in the testing set were those that had not been used in training, implying these examples were completely unseen to the classifiers. Each FER strategy was tested on the same testing set.

Each of the testing sequences was fed into each one of the four FER systems and their outputs were noted as correctly or incorrectly recognised according to the criterion for an accurately recognised facial expression.

Emotion	Label	Total Positive Examples Available	Positive Examples Used For Testing
Anger	1	45	31
Disgust	2	59	45
Fear	3	25	11
Happy	4	69	55
Sadness	5	28	14
Surprise	6	83	69
<b>Total</b>	–	<b>309</b>	<b>225</b>

TABLE 5.3: The testing data used in the FER accuracy experimentation.

### 5.2.3 Results and Analysis

A discussion and analysis of the results is carried out in the following subsections: Subsections 5.2.3.1–5.2.3.4 provide a detailed discussion on the results of each individual FER approach, followed by a comparison of the four methods in Subsection 5.2.3.5. A detailed table of the results obtained is provided in Table A.2 in Appendix A, with relevant excerpts of this table provided in the discussion in each subsection below. Also, confusion matrices of each method are provided in each relevant subsection below, but Appendix A provides all four confusion matrices in a single page in Tables A.3–A.6 should the reader wish to analyse them jointly.

#### 5.2.3.1 WFE Method Accuracy Results and Analysis

Emotion	Total Examples	WFE	
		Correct	Correct(%)
Anger	31	27	87
Disgust	45	38	84
Fear	11	6	54
Happy	55	50	90
Sadness	14	9	64
Surprise	69	57	82
<b>Overall</b>	<b>225</b>	<b>187</b>	<b>83.0</b>

TABLE 5.4: Facial expression recognition results of the *WFE* method.

Table 5.4 summarises the recognition accuracy of the *WFE* method for each emotion. Overall, the *WFE* method obtained a very high average recognition accuracy of 83.0% in recognising the six emotional expressions. To fully appreciate this result, it should be considered that the accuracy of using random guessing in place of the classifier to place each of the test sequences into one of six classes would be  $\frac{1}{6}$  which is approximately 16%. The classifier here performs at least five times better than random guessing. This is a

very encouraging result, especially considering that the number of training examples per emotion used to train the classifier was very limited.

Figure 5.5 graphically depicts the percentage accuracy of the *WFE* method for each emotion. It is observed that the emotion *Happy* obtained the highest accuracy of 90%, while the lowest, but by no means low, accuracy obtained was for *Fear* with an accuracy of 54%.

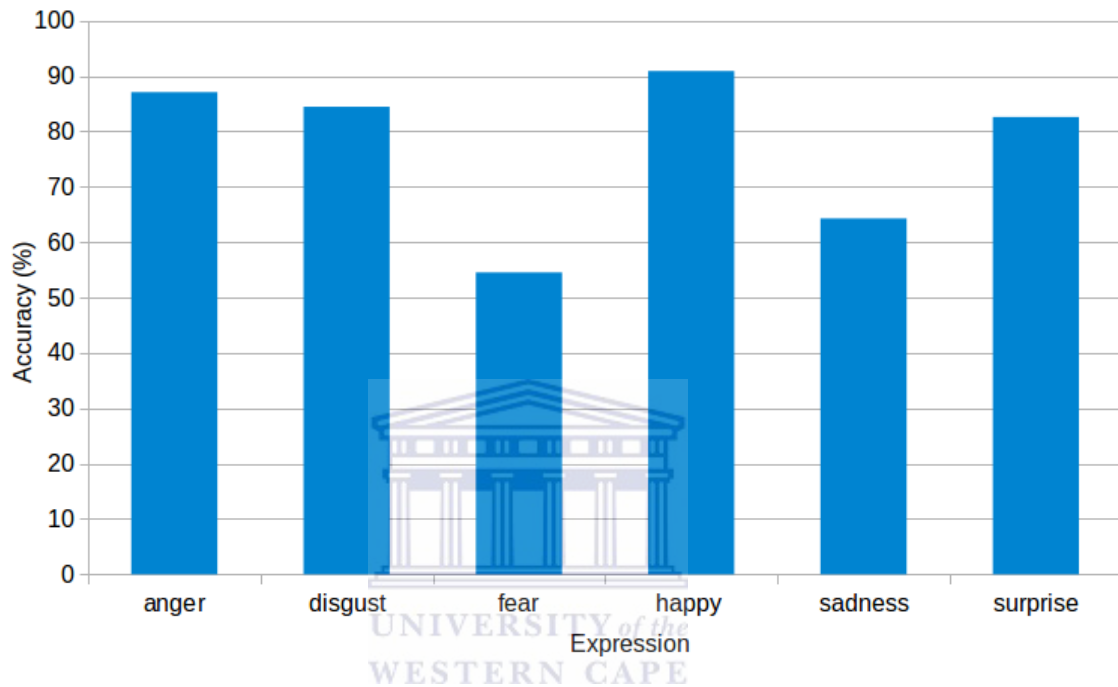


FIGURE 5.5: Recognition accuracy of the *WFE* method.

While 54% may appear to be a low accuracy, two facts should be considered. First, only 11 sequences of *Fear* were available for testing. Thus, with 6 of the 11 sequences misclassified, an apparently lower percentage accuracy results. Second, comparing this accuracy with the accuracy of random guessing reveals that it is still more than three times better.

It is also interesting to note that the two emotions with the lowest accuracy—*Fear* and *Sadness*, which achieved an accuracy of 64%—are also those with the fewest number of test sequences. Increasing the size of the training and testing data can be investigated in future as a means of improving the accuracy of recognising these emotions.

It is important to note that, aside from *Fear* and *Sadness*, all other emotions achieved accuracies of above 82%. The results obtained are very encouraging and it is clear that the motion-based feature extraction technique and segmentation procedure implemented in this research are effective and appropriate.

The high accuracy of *Happy* may be attributed to the fact that there is limited variation in the movement of the facial features of this emotion, where *Happy* is always linked to the raising of the cheeks and the dropping of the jaw, with the lips parted in the majority of cases.

<b>Actual</b>	<b>Predicted</b>						<b>Total</b>
	Anger	Disgust	Fear	Happy	Sadness	Surprise	
Anger	<b>27</b>	2	0	0	2	0	31
Disgust	2	<b>38</b>	0	0	5	0	45
Fear	0	0	<b>6</b>	2	3	0	11
Happy	0	0	0	<b>50</b>	5	0	55
Sadness	5	0	0	0	<b>9</b>	0	14
Surprise	0	0	1	0	11	<b>57</b>	69

TABLE 5.5: Confusion matrix of the *WFE* method accuracy results.

A confusion matrix of the accuracy results of the *WFE* method is provided in Table 5.5. For each emotion, the matrix specifies the distribution of system predictions across emotions. For example, the first row specifies that, of the 31 sequences of *Anger*, 27 were classified correctly as *Anger*, while two sequences were incorrectly classified as—and confused with—*Disgust* and another two, as *Sadness*.

In analysing the results of the confusion matrix, it is observed that the number of misclassified cases in the matrix is generally small and can be attributed to slight similarities between the movements of emotions in some sequences, the manner in which individual subjects perform these expressions and also to random factors caused by, among other things, limited data and classifier decisions.

The largest number of misclassified cases are of *Surprise* being confused with *Sadness*. This may be attributed to a similarity of the feature vectors of some test sequences of *Surprise* with those of *Sadness*. The emotion *Sadness* can be performed in many different ways. For example, *Sadness* could involve the raising of the cheeks and the tightening of the eye lids, but it could also involve the dropping of the jaw and parting of the lips, which are generally present in the emotion *Surprise*.

It is encouraging to note that, despite the relatively larger number of misclassified cases, this emotion (*Surprise*) achieves the highest accuracy. It is also important that the classifier is consistent in its decision and does not misclassify this emotion randomly.

The belief that *Sadness* may be performed in many different ways, making it similar to other emotions is strengthened by observing that the classifier confused a number of sequences of every emotion with *Sadness*. This indicates that this expression may

appear, in terms of the feature vector used, similar to other expressions in some of the test sequences.

Strangely *Sadness* was consistently misclassified as *Anger*. This disparity may be attributed to the specific sequences of each emotion whereby some sequences of every emotion appeared similar to *Sadness*, but the sequences of *Sadness* looked either like *Anger* or correctly as *Sadness*.

It is also very strange that *Happy* should be confused with its diametrical opposite *Sadness*, although this was in a small number of sequences. Again, this may be due to the manner in which this expression was performed in these test sequences, or to a classification decision by the classifier.

Ultimately, the reason for various classification decisions by the classifier are not clear and are difficult, if at all possible, to determine [31]. A further investigation in future with a larger amount of data may be more revealing.

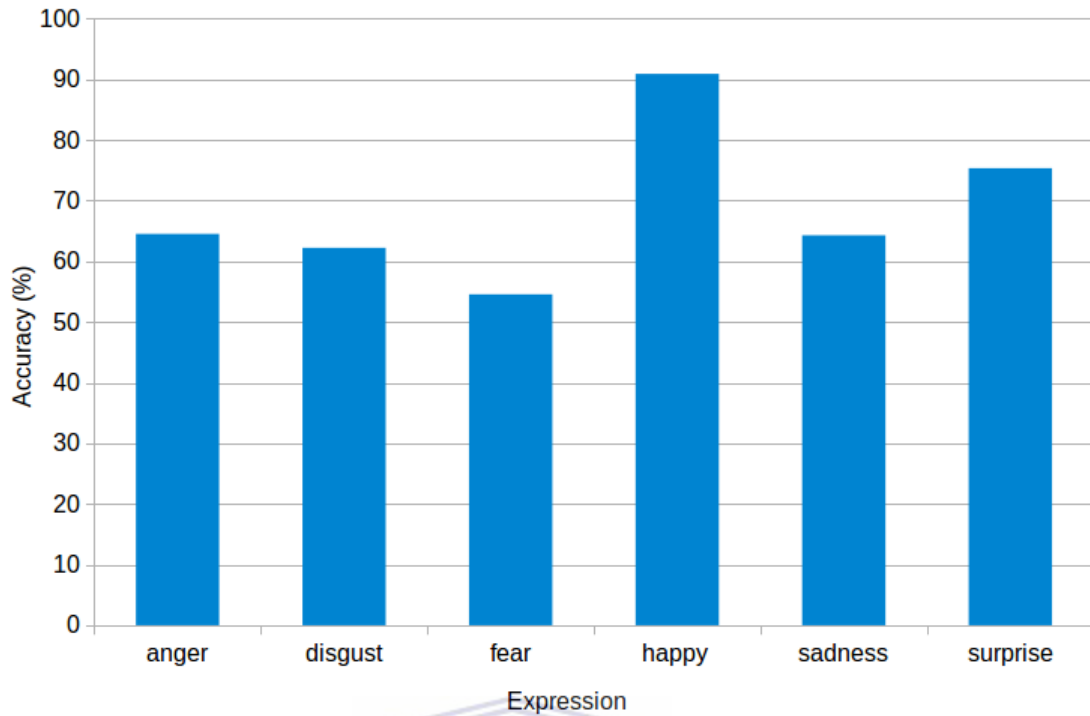
It can be concluded that, with an overall accuracy of 83%, the *WFE* strategy has no problem recognising the six basic emotional expressions.

### 5.2.3.2 AU Method Accuracy Results and Analysis

Emotion	Total	AU	
	Examples	Correct	Correct(%)
Anger	31	20	64
Disgust	45	28	62
Fear	11	6	54
Happy	55	50	90
Sadness	14	9	64
Surprise	69	52	75
<b>Overall</b>	<b>225</b>	<b>165</b>	<b>73.0</b>

TABLE 5.6: Facial expression recognition results of the *AU* method.

Table 5.6 is an excerpt of Table A.2 in Appendix A and summarises the recognition accuracy of the *AU* method for each emotion. Figure 5.6 graphically depicts a summary of the results per expression. It is seen that this method also obtained a high average recognition accuracy of 73.0%, although lower than the previous method. Once again, this accuracy should be taken in comparison to the random guessing accuracy in order to fully appreciate it. This is a very encouraging result and shows that recognising AUs and using them to infer the six basic emotional expressions is possible with a high accuracy.

FIGURE 5.6: Recognition accuracy of the *AU* method.

It is very interesting to note that the highest accuracy was once again achieved for *Happy* with the same accuracy of 90%, as is also with the lowest accuracy, achieved by *Fear* with the same accuracy of 54%. This observation strengthens the belief that the expression *Happy* experiences limited variation and leads to more consistent correct classification.

It can be seen from the figure that, aside from *Fear*, all other emotions achieved an accuracy of above 60% in recognising the six basic emotional expressions.

A confusion matrix for the AU recognition method is provided in Table 5.7.

<b>Actual</b>	<b>Predicted</b>						<b>Total</b>
	Anger	Disgust	Fear	Happy	Sadness	Surprise	
Anger	<b>20</b>	9	0	1	1	0	31
Disgust	14	<b>28</b>	0	0	3	0	45
Fear	1	0	<b>6</b>	2	0	2	11
Happy	0	2	0	<b>50</b>	3	0	55
Sadness	1	3	0	0	<b>9</b>	1	14
Surprise	3	2	3	2	7	<b>52</b>	69

TABLE 5.7: Confusion matrix of the *AU* method accuracy results.

It is seen from the matrix that while the number of misclassified sequences is still relatively low with this method, it is higher than that of the *WFE* method. It is interesting



to note, however, that confusion trends observed with the *WFE* method appear to generally hold with this method as well. For example, *Surprise* is mostly confused with *Sadness*, *Anger* with *Disgust* and *Happy* with *Sadness*. This indicates, impressively, that the *WFE* method classifier appears to consistently model the individually extracted AU features towards FER with greater success, despite the big difference in approach. It also confirms that appropriate features have been used in the system.

It is observed that, in this case, the largest number of misclassified sequences are for *Disgust* which was misclassified as *Anger* in 14 sequences. The production rules of these two expressions share AUs 17, 25 and 26, and may also share a unique facial movement involving the wrinkling of the nose, depending on how they are performed. This could be the reason for the relatively large number of misclassified cases with this specific method. The link between the two expressions is further evidenced by the fact that the second highest number of misclassified sequences--9 sequences--is for *Anger* misclassified as *Disgust*. The previous method made use of a classifier can model the difference between these expressions more effectively.

As with the *WFE* method, almost every emotion appears to be misclassified as *Sadness* in some sequences. Unlike the *WFE* method, *Surprise* is far less consistently misclassified in this method, having been misclassified as every other expression in some sequences. This can most likely be attributed to incorrect classifications by the AU classifiers relevant to this emotion in a number of sequences of this emotion, the result of which was an incorrect assignment to a different emotion in each case. Nevertheless, this emotion still achieves a very high accuracy of 75%.

It can be concluded that this method is also effective for FER, with an overall accuracy of 73%.

### 5.2.3.3 HybridWFEFirst Method Accuracy Results and Analysis

Emotion	Total Examples	HybridWFEFirst	
		Correct	Correct(%)
Anger	31	27	87
Disgust	45	39	86
Fear	11	8	72
Happy	55	51	92
Sadness	14	9	64
Surprise	69	57	82
<b>Overall</b>	<b>225</b>	<b>191</b>	<b>84.5</b>

TABLE 5.8: Facial expression recognition results of the *HybridWFEFirst* method.

Table 5.8 is an excerpt of Table A.2 in Appendix A and summarises the recognition accuracy of the *HybridWFEFirst* method for each emotion. Figure 5.7 graphically depicts a summary of the results per expression for convenience.

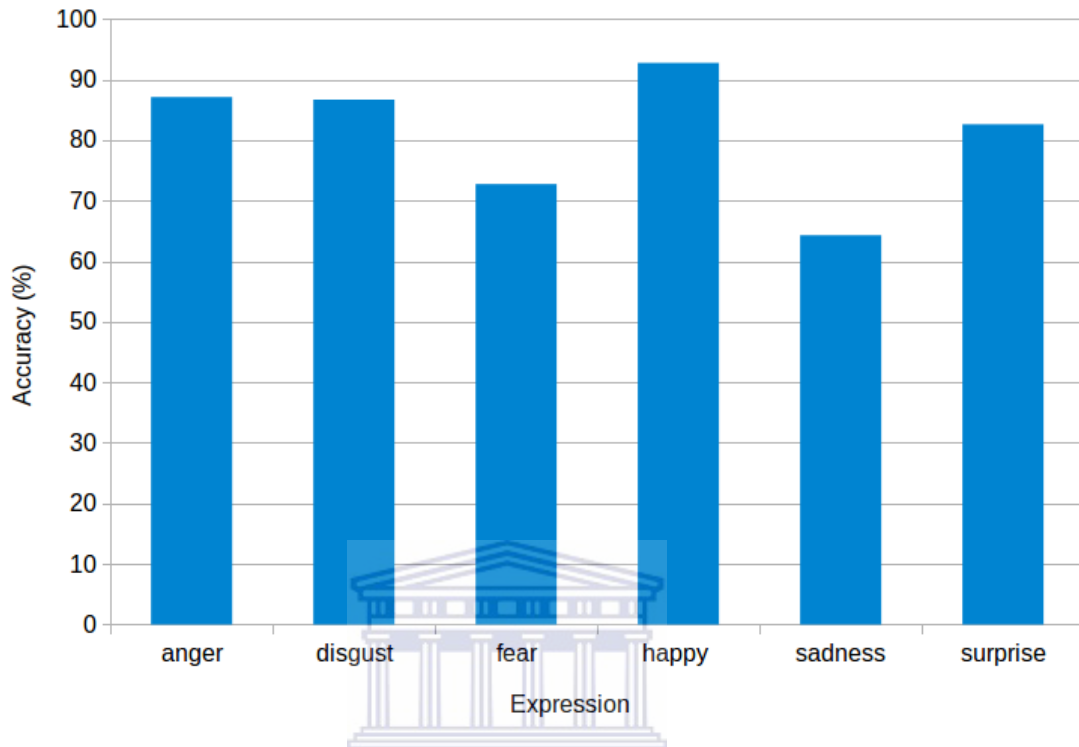


FIGURE 5.7: Recognition accuracy of the *HybridWFEFirst* method.

This method, which is the first hybrid method to be analysed, is observed to obtain a very high average recognition accuracy of 84.5%, higher than both previous methods. As before, this accuracy is several times higher than the random guessing accuracy. This result is extremely encouraging and indicates a successful hybrid implementation.

The individual emotion accuracies range from 64% to 92%. It is seen that, yet again, the emotion *Happy* obtains the highest accuracy for this method. The previous assertion that this emotion is very elaborate and unique is further strengthened.

Contrary to the previous two methods, however, the lowest—but not low—accuracy obtained is for *Sadness* in this method, although *Fear* still achieves the second lowest accuracy of 72%. The minimum and maximum accuracies for this method are, respectively, much higher and slightly higher, than the previous two methods.

It is important to note that, aside from *Sadness*, all other emotions achieved an accuracy of above 70% in recognising the six basic emotional expressions. Also, four of the emotional expressions achieved very high recognition accuracies of above 80% and three expressions achieved extremely high accuracies of above 85%.

These results are very encouraging and it is clear that combining the *AU* and the *WFE* methods in this way produces excellent results. It also confirms the statement made in Chapter 4 that this hybrid is expected to perform at least as well as the *AU* method.

A confusion matrix for this method is provided in Table 5.9. Referring to the matrix, it is encouraging to note that the number of misclassified cases is small.

<b>Actual</b>	<b>Predicted</b>						<b>Total</b>
	Anger	Disgust	Fear	Happy	Sadness	Surprise	
Anger	<b>27</b>	3	0	0	1	0	31
Disgust	3	<b>39</b>	0	0	3	0	45
Fear	1	0	<b>8</b>	2	0	0	11
Happy	0	0	0	<b>51</b>	4	0	55
Sadness	5	0	0	0	<b>9</b>	0	14
Surprise	0	0	1	3	8	<b>57</b>	69

TABLE 5.9: Confusion matrix of the *HybridWFEFirst* method accuracy results.

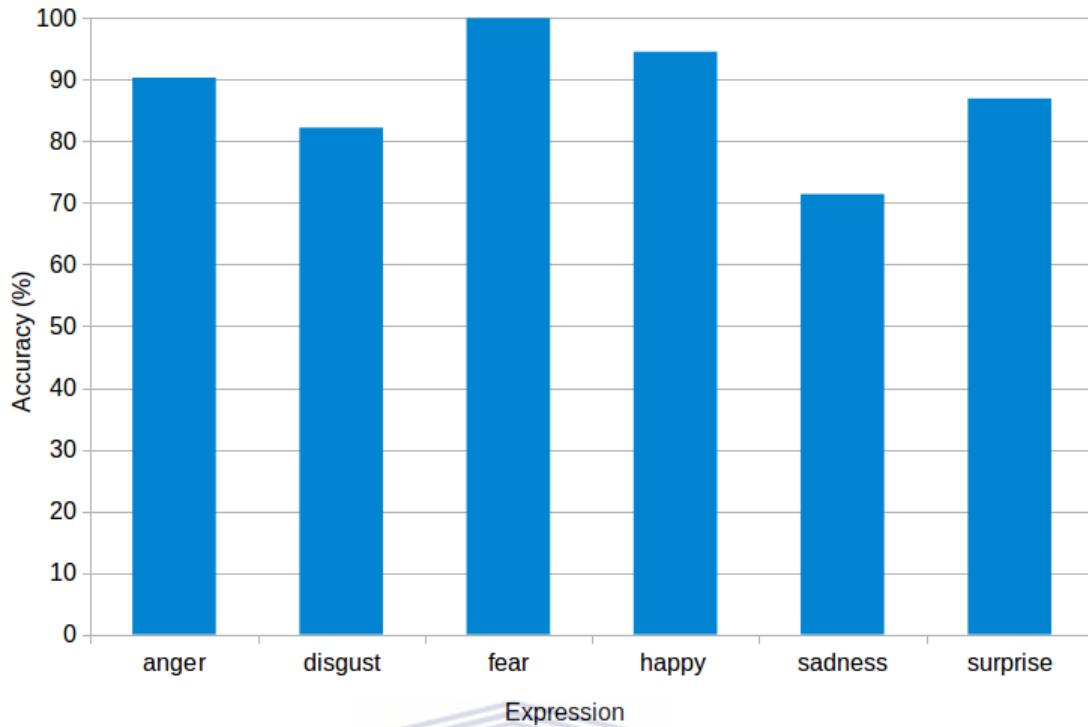
It is observed that the lowest performing expression *Sadness* is consistently misclassified as *Anger*, and a number of sequences of almost every emotion are confused with the former expression. It is interesting to note that very similar trends were also observed for the *WFE* and *AU* methods. This further strengthens the belief that some test sequences of each emotion may appear as *Sadness*, which may be performed in many different ways, but that the sequences of *Sadness* were performed like *Anger* or correctly as *Sadness*. Also, as with previous methods, *Surprise* is confused with *Sadness* in the majority of its cases.

It is quite clear that this hybrid method performs very well, with an obtained average recognition accuracy of 84.5%. This can be due to the fact that it takes the best qualities from the *WFE* and *AU* methods, combining small-scale facial features and movements on a global scale.

#### 5.2.3.4 HybridAUFIRST Method Accuracy Results and Analysis

Table 5.10 is an excerpt of Table A.2 in Appendix A and summarises the recognition accuracy of the *HybridAUFIRST* method for each emotion. Figure 5.8 graphically depicts a summary of the results per expression for convenience.

Referring to Table 5.10, it is seen that this method obtained a very high average recognition accuracy of 88%. The accuracies range from a very high minimum accuracy of 71% to a perfect accuracy of 100%. With this method, even the lowest accuracy emotion is predicted at an accuracy approximately 4.5 times better than random guessing. This encouraging result indicates a very successful hybrid implementation.

FIGURE 5.8: Recognition accuracy of the *HybridAUFIRST* method.

Emotion	Total Examples	HybridAUFIRST	
		Correct	Correct(%)
Anger	31	28	90
Disgust	45	37	82
Fear	11	11	100
Happy	55	52	94
Sadness	14	10	71
Surprise	69	60	86
<b>Overall</b>	<b>225</b>	<b>198</b>	<b>88.0</b>

TABLE 5.10: Facial expression recognition results of the *HybridAUFIRST* method.

For this method, the emotion *Fear* obtained the highest accuracy of 100%, and this was closely followed by *Happy* with 94%, which consistently achieved the highest accuracy in other methods. It is encouraging to note that *Fear* that was the lowest, or among the lowest, performing emotions in previous methods springs up with a perfect accuracy in this method. In the absence of the emotion *Fear*, the lowest—but by no means low—accuracy with this method belongs to *Sadness*, which was the second or third lowest accuracy emotion in other methods.

It is encouraging to say that, for this method, all classifiers achieve above 70% accuracies. Aside from *Sadness*, however, all other emotions achieve accuracies of above 80%, with three out of the six emotions achieving outstanding accuracies of above 90%. This is

also the only classifier that achieves a perfect accuracy of 100% for an emotion.

<b>Actual</b>	<b>Predicted</b>						<b>Total</b>
	Anger	Disgust	Fear	Happy	Sadness	Surprise	
Anger	<b>28</b>	2	0	0	1	0	31
Disgust	6	<b>37</b>	0	0	2	0	45
Fear	0	0	<b>11</b>	0	0	0	11
Happy	3	0	0	<b>52</b>	0	0	55
Sadness	3	0	1	0	<b>10</b>	0	14
Surprise	0	0	0	5	4	<b>60</b>	69

TABLE 5.11: Confusion matrix of the *HybridAUFIRST* method accuracy results.

A confusion matrix of the results of this method are provided in Table 5.11.

Analysing the matrix reveals very similar trends to those observed in the confusion matrices of previous methods. It is once again seen that *Surprise* is confused with *Sadness* in a large number (but, in this case, not the majority) of cases. *Sadness* is confused with *Anger* in the majority of cases. On the other hand, it is observed that far fewer sequences of each emotion were confused with *Sadness* in this method, which indicates a more consistent classification outcome.

One interesting observation is that sequences of *Fear*, that were largely confused with *Disgust* and *Sadness* in other methods, are consistently classified correctly with this method.

Also, whereas *Happy* was mostly confused with *Sadness* in other methods, it is confused with *Anger* in this method, although only in a very small number of cases.

It is evident that this hybrid method is a very effective FER strategy, having achieved an outstanding overall accuracy of 88.0%. This shows that the probability of AUs being present using these as features for a classifiers can be very effective at recognising the six basic emotional expressions.

It is concluded at this point that the first research question

### 5.2.3.5 Comparison of Methods

This section compares the results the four methods based on the results obtained in the previous section, aiming to answer the second research question posed in Chapter 1: “How do the hybrid approaches compare with traditional whole FER approaches in terms of FER accuracy?”.

A comparative graphical depiction of the average accuracy of each method across all six emotions is provided in Figure 5.9. The figure, as well as Figure 5.10 discussed later, are based on the complete set of results provided in Table 5.4 in Appendix A.

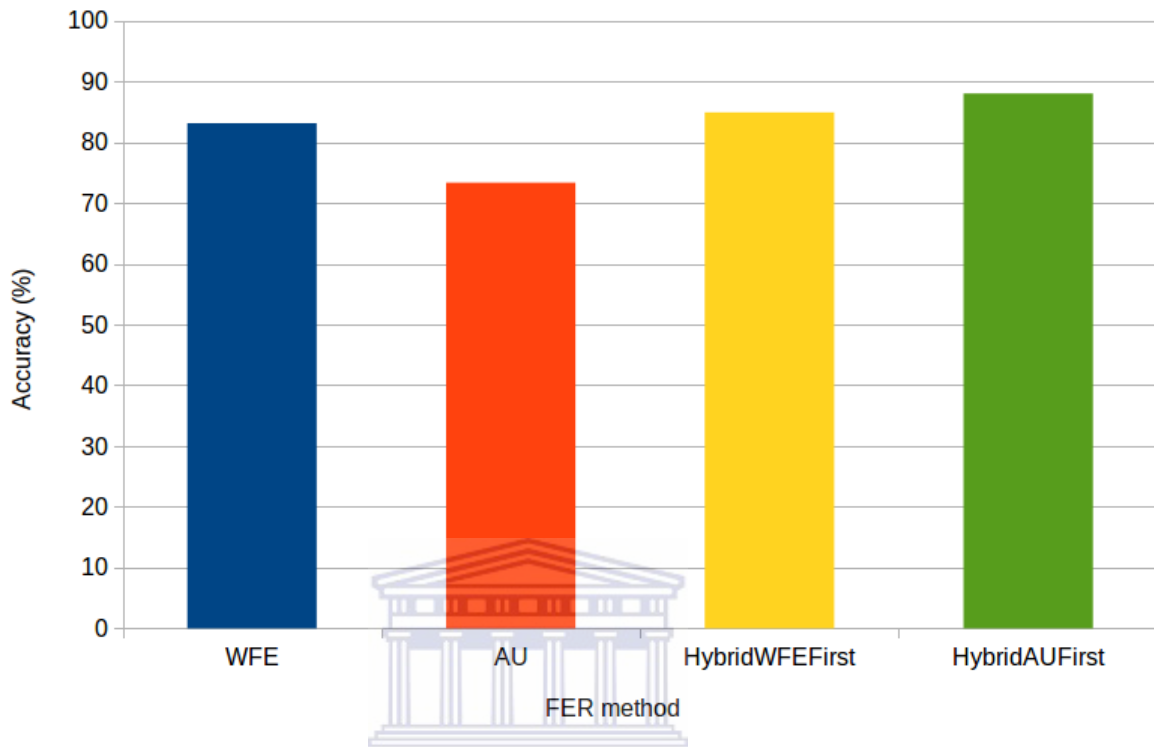


FIGURE 5.9: A graphical depiction of the average accuracy of each method across all six emotions.

Referring to Figure 5.9, it is very encouraging to note that, overall, all four methods achieve very high accuracies of above 70%. It can be seen that the *HybridAUFIRST* method has the highest accuracy, higher than its immediate successor—*HybridWFEFirst*—by 5.0%. *HybridWFEFirst*, in turn, performs at a higher accuracy than the *WFE* method by a smaller margin of 1.5%. Finally, the *WFE* method outperforms the *AU* method by a large margin of 10.0%. The difference between the highest performing method—*HybridAUFIRST*—and the lowest performing method—the *AU* method—is a very large margin of 15.0%.

Each of these margins, however big or small, can contribute significantly to a final system deployed in the real-world, provided they are statistically significant. As the number of processed sequences increases, seemingly small margins in accuracy can lead to increasing differences in accuracy. For example, a system that processes 100 million sequences can benefit from even a 1% increase in accuracy, if statistically significant, whereby one million sequences would be prevented from being incorrectly classified under such a seemingly small increase. Therefore, any increase in accuracy is very advantageous during deployment, provided it is statistically significant.

In order to determine whether the differences in the overall accuracies in the methods were statistically significant, McNemar’s test was applied to results of each method. McNemar’s test is a pair-wise version of the Chi-Square test in which the degrees of freedom are always 1. Each pair of methods was compared resulting in a total of six comparisons. The tables used in these comparisons are provided in Tables A.7–A.12 in Appendix A, and the resulting Chi-Square and  $p$ -values are provided in Table 5.12.

Method 1	Method 2	Chi-Square	$p$ -value
WFE	AU	11.025	0.0009
WFE	HybridWFEFirst	1.125	0.2888
WFE	HybridAUFfirst	9.091	0.0026
AU	HybridWFEFirst	22.321	< 0.0001
AU	HybridAUFfirst	26.256	< 0.0001
HybridWFEFirst	HybridAUFfirst	5.143	0.0233

TABLE 5.12: Results of McNemar’s test for comparing the four FER methods.

Analysing the results in Table 5.12, it is becomes apparent that:

- The difference in accuracy between the *AU* method and every other method is extremely significant.
- The difference in accuracy between the *WFE* method and *HybridWFEFirst* method is slightly significant, if at all. The two methods perform very similarly in terms of accuracy.
- The difference in accuracy between the *HybridAUFfirst* and the *WFE* and *AU* methods is extremely significant, that of *HybridAUFfirst* and *HybridWFEFirst* is very significant.

From these results, it becomes apparent that the hybrid implementations both perform significantly better than the (traditional) *AU* method, and *HybridAUFfirst* performs significantly better than the (traditional) *WFE* method. In a real-world implementation, these accuracies can prove to be very valuable.

The results of each FER method per expression, as percentages of the total test sequences, can be viewed in Figure 5.10.

In confirmation of the previous findings, it is clear, when looking at the figure, that the *HybridAUFfirst* method outperforms all other strategies in recognising all but one emotional expression, namely, *Disgust*. It is also clear that the *WFE* and *HybridWFE-First* method appear to perform at a very similar level for all the emotions except *Fear*, with *HybridWFEFirst* appearing to achieve very slightly higher accuracies in some cases.

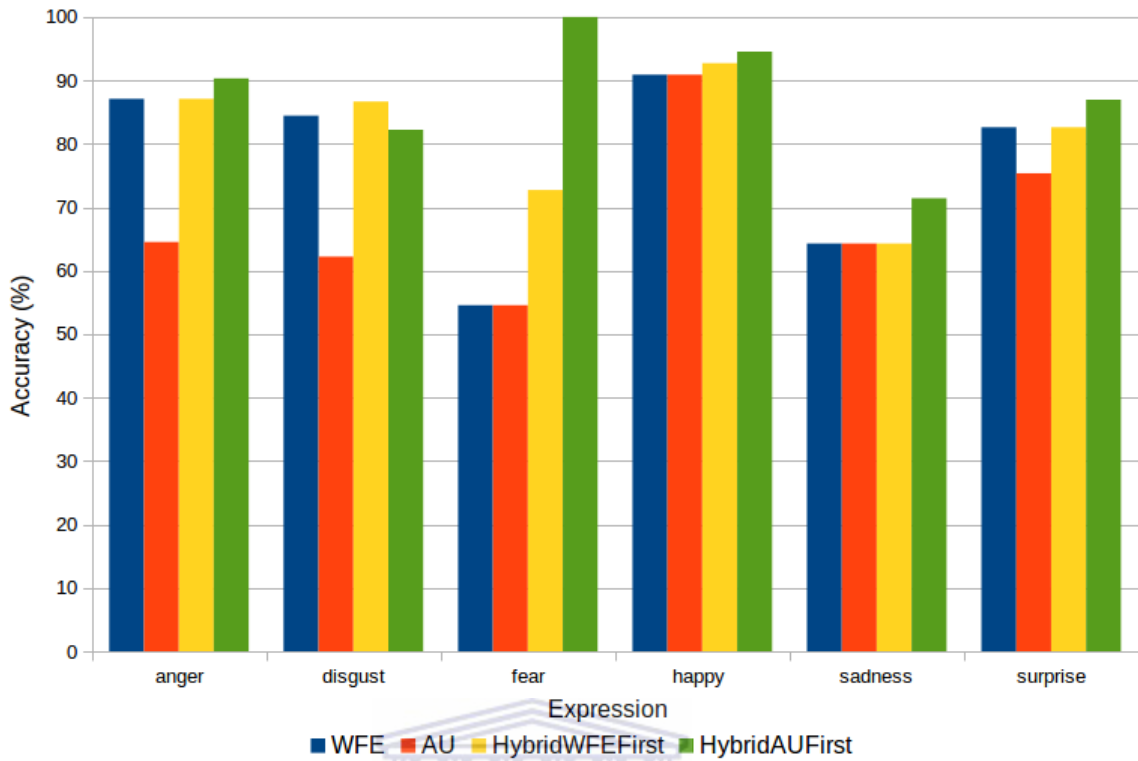


FIGURE 5.10: A graphical depiction of the accuracy of each FER method per expression.

All the methods perform at least as well as the *AU* method in all the emotions, but substantially better in many. All of these points stand to confirm the previous findings.

When looking at the emotion *Fear*, it can be seen that the *AU* method ties with the *WFE* method, and when looking at the emotion *Sadness*, it is seen that the *AU* method ties with both the *WFE* and *HybridWFEFirst* method. A similar case is observed for *Happy*, with the hybrid approaches achieving small increases in accuracy over the *WFE* and *AU* methods, and in this case all the methods achieved exceptionally high accuracies.

The most impressive feature of the *HybridAUFIRST* method is its consistency in recognizing the six basic emotional expressions at a high accuracy, as the accuracy of no emotion falls below the 70% mark. In contrast, the accuracies of other methods fall to below 70%, with the *AU* and *WFE* methods even falling below 60% accuracy although, as previously noted, an accuracy of 50% should by no means be considered a low accuracy.

It appears that the use of AUs can help improve FER accuracy, but this improvement is best realized when a classifier is used to learn varied AU presence levels and associate them with emotions as in *HybridAUFIRST*, as opposed to using the set of pre-determined production rules as in *HybridWFEFirst* and the *AU* method. This is attributed to the



fact that the production rules, while still yielding high accuracies, may be less effective at modelling variations in performance of the six basic emotional expressions. The classifier, on the other hand, can dynamically learn to recognise these variations.

At this stage, it can be said that the final research objective 5, which required that a comparison of the hybrid and traditional FER approaches be carried out, has been successfully achieved. Accordingly, in response to research question 2 set out in Chapter 1 which was phrased as “How do the hybrid approaches compare with traditional whole FER approaches in terms of FER accuracy?”, it is stated that the *HybridAUFIRST* hybrid approach performs significantly better, in terms of FER accuracy, than the traditional approaches and the *HybridWFEFIRST* hybrid approach performs significantly better than the *AU* traditional method, and at the same level as the *WFE* traditional method.

### 5.3 Summary and Conclusions

This chapter discussed the two experiments carried out to answer the two research questions set out in Chapter 1. The first experiment involved determining the accuracy of the 16 AU classifiers used in the *AU*, *HybridWFEFIRST* and *HybridAUFIRST* methods, thereby comparing the use of local and global segmentation in each AU classifier. The second experiment aimed to determine and compare the FER accuracy of the four proposed FER methods in order to determine how hybrid methods compare with the traditional methods first described in Chapter 1.

The results of the AU experimentation were analysed and it became clear that, for all 16 AUs, both segmentation methods were highly accurate. The global segmentation procedure accuracy was compared to that of the local segmentation procedure accuracy. It was found that global segmentation was more suited to some AUs—8 of the 16 AUs—while local segmentation yielded a higher accuracy for other AUs—6 AUs—and in still others—2 AUs—it the accuracy was the same under both types of segmentation. It was concluded that a choice of segmentation technique depends on the specific AU to be recognised.

A detailed analysis of the FER accuracy of the four proposed methods was also carried out and revealed that all four techniques are capable of obtaining very high accuracies. The *HybridAUFIRST* method achieved the highest accuracy out of all four techniques, followed by *HybridWFEFIRST*, the *WFE* method, and finally the *AU method*. A statistical test revealed that the small difference in accuracy between the *HybridWFEFIRST* and *WFE* method—with the *HybridWFEFIRST* being higher—was not significant, but that

the difference in accuracy of *HybridAUFIRST* with every other method was very significant. It was concluded that the *HybridAUFIRST* hybrid approach is more accurate than the traditional methods, and *HybridWFEFIRST* is more accurate than the *AU* traditional method, but only as good as the *WFE* traditional method.

The next chapter concludes the thesis.



## Chapter 6

# Conclusion

This research aimed at creating fully automatic robust facial expression recognition systems that utilises the Facial Action Coding System for the purpose of recognising six whole facial expressions namely: anger, happy, disgust, fear, sadness and surprise. This research also aimed at doing a comparison between four unique FER approaches that recognise whole facial expressions. Two of the FER approaches are traditional approaches while the other two are hybrid approaches. All four FER approaches utilised dense flow for feature extraction and Support Vector machines for classification of the features.

The first approach referred to in the research as the *WFE* method makes use of a multi-class SVM that is directly trained on the features extracted using dense flow. The multi-class SVM is trained to recognise the six whole facial expressions. The second approach referred to in this research as the *AU* method uses a two-step procedure in which 16 relevant AUs are first recognised by means of 16 individual SVMs, each trained to recognise a specific AU, followed by the application of a set of production rules on the presence or absence of the AUs to carry out FER.

The third approach referred to in this research as the *HybridWFEFirst* method combines the *WFE* and *AU* methods such that the prediction of the *WFE* method serves as an initial FER prediction which is then either confirmed or corrected by the *AU* method. The fourth and final approach referred to in this research as the *HybridAUFIRST* method also combines subsets of the *WFE* and *AU* methods such that the output of the 16 individual SVM classifiers from the *AU* method are then used as features in a multi-class SVM similar to the one used in the *WFE* method to recognise the six whole facial expressions.

In response to research question 1 which asked, Can robust autonomous hybrid FER systems be created utilising the FACS towards recognition of WFEs?, it was shown that not only can such systems be created but they can also achieve brilliant results in recognising six whole facial expressions. Two hybrid systems were created *HybridAUFfirst* and *HybridWFEFirst* which were derived from two traditional approaches *WFE* and *AU*.

In response to research research question 2 which asked, How do the hybrid approaches compare with traditional whole FER approaches in terms of FER accuracy?, it is stated that the *HybridAUFfirst* hybrid approach performs significantly better, in terms of FER accuracy, than the traditional approaches with an obtained average accuracy of 88% and the *HybridWFEFirst* hybrid approach performs significantly better than the *AU* traditional method, and at the same level as the *WFE* traditional method with an obtained average accuracy of 81%

In response to the final research question 3 which asked How does the use of local and global segmentation of the face during feature extraction compare towards AU recognition accuracy?, it is concluded that both feature extraction techniques result in very high accuracy in AU recognition, but either technique may be somewhat more suited to some AUs than others, while there is no difference for some AUs. The choice is AU-specific.

These finding have made a significant contribution to the field of facial expression recognition. It has provided a comparative study between two traditional FER approaches *AU* and *WFE* methods that has, in the past been a hotly contested subject as to which approach performs the best. The findings have also made it clear that hybrid approaches *HybridWFEFirst* and *HybridAUFfirst* methods combining both *AU* and *WFE* methods produce a much better average recognition accuracy.

This research has also significantly contributed to the SASL group in that it produced an improved FER approach *HybridAUFfirst* method obtaining an average accuracy of 88%. The improved FER approach can thus be integrated in the SASL gesture recognition system to improve the accuracy in recognising sign language gestures.

## 6.1 Future Work

This research only focused on segmenting the face as part of the preprocessing before feature extraction was implemented. Normalising the face by way of an affine or perspective transformation could improve the accuracy as facial expressions often result in tilting of the face causing an in plane rotation. A normalization procedure would negate

in plane rotation and thus adding to the robustness and accuracy of the system. This research focused solely on frontal face images. It would be possible to extend the research to recognise expressions at different degrees of rotation and thereby improving the robustness of the system.

In this research the experimentation was done using the CK and CK+ database. None of the databases included SASL or any other kind of sign language facial expressions. Furthermore, no database containing facial expressions in SASL exists. Seeing as this research is done in the context of SASL gesture recognition it becomes necessary to thoroughly test the effectiveness of the systems when a database containing SASL facial expressions exist. It's important to note that given how expressive sign language facial expressions are an improvement in the accuracy of the system is very likely if experimentation is done.

In this research three different ROIs were produced from the three different segmentation procedures namely: upper, lower and whole face. All the ROIs were set to the same size before the extraction technique was applied. Experimenting with optimising the sizes of the ROIs could produce improved results and thus warrants inclusion into future work.

## 6.2 Concluding Remarks

Through the duration of this research, the researcher has gained a huge amount of knowledge and experience in the field of computer vision and more specifically FER. It is a hope that this research will add value to the work of other researchers specialising in FER and ultimately contribute to the betterment of the ever growing knowledge-base that is the computer vision community. It is also a hope that this research will add significant value towards the advancements of the SASL project.

# Appendix A

## Additional Test Results

AU Recog.	Total Test Examples	Global Segm.		Local Segm.	
		Correct	Correct (%)	Correct	Correct (%)
1	112	105	93	100	89
2	72	66	91	66	91
4	116	93	80	91	78
5	52	42	80	41	78
6	78	69	88	65	83
7	76	57	75	57	75
9	40	34	85	35	87
10	16	10	62	11	68
12	72	60	83	61	84
15	52	44	84	41	78
16	18	13	72	12	66
17	110	97	88	91	82
20	48	39	81	41	85
23	32	21	65	25	78
25	152	134	88	132	86
26	28	23	82	24	85

TABLE A.1: AU classifier recognition accuracy results.

Emotion	Total Test Examples	WFE		AU		Hybrid WFEFirst		Hybrid AUFfirst	
		Cor.	Cor.(%)	Cor.	Cor.(%)	Cor.	Cor.(%)	Cor.	Cor.(%)
Anger	31	27	87	20	64	27	87	28	90
Disgust	45	38	84	28	62	39	86	37	82
Fear	11	6	54	6	54	8	72	11	100
Happy	55	50	90	50	90	51	92	52	94
Sadness	14	9	64	9	64	9	64	10	71
Surprise	69	57	82	52	75	57	82	60	86
<b>Overall</b>	<b>225</b>	<b>187</b>	<b>83.0</b>	<b>165</b>	<b>73.0</b>	<b>191</b>	<b>84.5</b>	<b>198</b>	<b>88.0</b>

TABLE A.2: Facial expression recognition results of the four methods: AU, WFE, HybridWFEFirst and HybridAUFfirst.

Actual	Predicted						Total
	Anger	Disgust	Fear	Happy	Sadness	Surprise	
Anger	<b>27</b>	2	0	0	2	0	31
Disgust	2	<b>38</b>	0	0	5	0	45
Fear	0	0	<b>6</b>	2	3	0	11
Happy	0	0	0	<b>50</b>	5	0	55
Sadness	5	0	0	0	<b>9</b>	0	14
Surprise	0	0	1	0	11	<b>57</b>	69

TABLE A.3: Confusion matrix of the WFE method accuracy results.

Actual	Predicted						Total
	Anger	Disgust	Fear	Happy	Sadness	Surprise	
Anger	<b>20</b>	9	0	1	1	0	31
Disgust	14	<b>28</b>	0	0	3	0	45
Fear	1	0	<b>6</b>	2	0	2	11
Happy	0	2	0	<b>50</b>	3	0	55
Sadness	1	3	0	0	<b>9</b>	1	14
Surprise	3	2	3	2	7	<b>52</b>	69

TABLE A.4: Confusion matrix of the AU method accuracy results.

Actual	Predicted						Total
	Anger	Disgust	Fear	Happy	Sadness	Surprise	
Anger	<b>27</b>	3	0	0	1	0	31
Disgust	3	<b>39</b>	0	0	3	0	45
Fear	1	0	<b>8</b>	2	0	0	11
Happy	0	0	0	<b>51</b>	4	0	55
Sadness	5	0	0	0	<b>9</b>	0	14
Surprise	0	0	1	3	8	<b>57</b>	69

TABLE A.5: Confusion matrix of the HybridWFEFirst method accuracy results.

Actual	Predicted						Total
	Anger	Disgust	Fear	Happy	Sadness	Surprise	
Anger	<b>28</b>	2	0	0	1	0	31
Disgust	6	<b>37</b>	0	0	2	0	45
Fear	0	0	<b>11</b>	0	0	0	11
Happy	3	0	0	<b>52</b>	0	0	55
Sadness	3	0	1	0	<b>10</b>	0	14
Surprise	0	0	0	5	4	<b>60</b>	69

TABLE A.6: Confusion matrix of the HybridAUFIRST method accuracy results.



		AU		Total
		1	0	
WFE	1	156	31	187
	0	9	29	38
Total		165	60	225

TABLE A.7: McNemar's Test for the WFE and AU methods.

		HybridWFEFirst		Total
		1	0	
WFE	1	166	21	187
	0	25	13	38
Total		191	34	225

TABLE A.8: McNemar's Test for the WFE and HybridWFEFirst methods.

		HybridAUFIRST		Total
		1	0	
WFE	1	187	0	187
	0	11	27	38
Total		198	27	225

TABLE A.9: McNemar's Test for the WFE and HybridAUFIRST methods.

		HybridWFEFirst		Total
		1	0	
AU	1	164	1	165
	0	27	33	60
Total		191	34	225

TABLE A.10: McNemar's Test for the AU and HybridWFEFirst methods.

		HybridAUFIRST		Total
		1	0	
AU	1	162	3	165
	0	36	24	60
Total		198	27	225

TABLE A.11: McNemar's Test for the AU and HybridAUFIRST methods.

	HybridAUFIRST		Total
	1	0	
Hybrid - WFE - First	1	191	191
	0	7	34
Total	198	27	225

TABLE A.12: McNemar's Test for the HybridWFEFirst and HybridAUFIRST methods.

# Bibliography

- [1] 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.
- [2] I. Achmed and J. Connan, “Upper body pose estimation towards the translation of south african sign language,” in *Proceedings of the Southern Africa Telecommunication Networks and Applications Conference*, 2010, pp. 427–432.
- [3] I. Achmed, I. M. Venter, and P. Eisert, “A framework for independent hand tracking in unconstrained environments,” in *Proceedings of the South African Telecommunication Network and Applications Conference SATNAC 2012*, 2012, pp. 159–164.
- [4] T. Borovicka, M. Jirina Jr, P. Kordik, and M. Jirina, “Selecting representative data sets,” *Advances in Data Mining Knowledge Discovery and Applications. Intech*, 2012.
- [5] C. A. Bouman, “Digital image processing : Connected component analysis,” January 2014, [Online] Available at <https://engineering.purdue.edu/~bouman/ece637/notes/pdf/ConnectComp.pdf>.
- [6] G. Bradski and A. Kaehler, *Learning OpenCV*. O’Reilly Media, 2008.
- [7] D. Brown, “Upper body pose recognition and estimation towards the translation of South African Sign Language,” Master’s thesis, University of the Western Cape, Computer Science, 2013.
- [8] D. L. Brown, M. Ghaziasgar, and J. Connan, “Faster upper body pose estimation using CUDA,” in *Proc. Southern Africa Telecommunication Networks and Applications Conference*, 2012.
- [9] G. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, A. B. Rosen *et al.*, “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.

- [10] 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.
- [11] J. F. Cohn, A. J. Zlochower, J. 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.
- [12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [13] 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.
- [14] Edmark, "Fripple place," January 2015, [Online] Available at <http://www.riverdeep.net/edconnect/softwareactivities/criticalthinking/frippleplace.jhtml>.
- [15] P. Ekman and W. Friesen, "The Facial Action Coding System: A technique for the measurement of facial movement," 1978.
- [16] P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique." 1994.
- [17] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.
- [18] P. Ekman and W. V. Friesen, "Facial Action Coding System," 1977.
- [19] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of non-verbal behavior*, vol. 6, no. 4, pp. 238–252, 1982.
- [20] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.
- [21] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.
- [22] I. Essa, A. P. Pentland *et al.*, "Coding, analysis, interpretation, and recognition of facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 757–763, 1997.

- [23] I. A. Essa, “Analysis, interpretation and synthesis of facial expressions,” Ph.D. dissertation, Citeseer, 1994.
- [24] G. Farneback, “Two-frame motion estimation based on polynomial expansion,” in *Image Analysis*. Springer, 2003, pp. 363–370.
- [25] R. Foster, “A comparison of machine learning techniques for hand shape recognition,” Master’s thesis, University of the Western Cape, Computer Science, 2014.
- [26] R. Foster, M. Ghaziasgar, J. Connan, and R. Dodds, “A comparison of machine learning techniques for hand shape recognition,” in *South African Telecommunication Networks and Applications Conference*, 2014, pp. 173–178.
- [27] I. Frieslaar, “Robust south african sign language gesture recognition using hand motion and shape,” Master’s thesis, University of the Western Cape, Computer Science, 2014.
- [28] I. Frieslaar, M. Ghaziasgar, and J. Connan, “Addressing the problem of hand occlusion in bimanual hand shape recognition,” in *South African Telecommunication Networks and Applications Conference*, 2014, pp. 241–246.
- [29] M. D. Gall, W. R. Borg, and J. P. Gall, *Educational research: An introduction*. Longman Publishing, 1996.
- [30] M. Ghaziasgar and J. Connan, “Investigating the feasibility factors of synthetic sign language visualization methods on mobile phones,” in *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*. ACM, 2010, pp. 86–92.
- [31] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.
- [32] B. K. Horn and B. G. Schunck, “Determining optical flow,” in *1981 Technical symposium east*. International Society for Optics and Photonics, 1981, pp. 319–331.
- [33] S. Howard, “Finger talk-south african sign language dictionary,” *South Africa: Mondri*, 2008.
- [34] C. Hsu, C. Chang, and C. Lin, “A practical guide to support vector classification,” National Taiwan University, Tech. Rep., 2003.
- [35] C. E. Izard, “Innate and universal facial expressions: evidence from developmental and cross-cultural research.” 1994.

- [36] 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.
- [37] A. Kapoor, "Automatic facial action analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.
- [38] M. Kenji, "Recognition of facial expression from optical flow," *IEICE TRANSACTIONS on Information and Systems*, vol. 74, no. 10, pp. 3474–3483, 1991.
- [39] B. V. Kumar, "Face expression recognition and analysis: the state of the art," *Course Paper, Visual Interfaces to Computer*, 2009.
- [40] P. Le Bek, "Learning to recognise actions in egocentric video," Master's thesis, University Glasgow, School of Computing Science, 2014.
- [41] P. Li, "Hand shape estimation for South African Sign Language," Master's thesis, University of the Western Cape, Computer Science, 2010.
- [42] P. Li, M. Ghaziasgar, and J. Connan, "Hand shape recognition and estimation for south african sign language," in *South African Telecommunication Networks and Applications Conference*, 2011, pp. 344–349.
- [43] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.
- [44] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li, "Detection, tracking, and classification of action units in facial expression," *Robotics and Autonomous Systems*, vol. 31, no. 3, pp. 131–146, 2000.
- [45] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision." in *IJCAI*, vol. 81, 1981, pp. 674–679.
- [46] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.
- [47] P. Martins, J. Sampaio, and J. Batista, "Facial expression recognition using active appearance models." in *VISAPP (2)*, 2008, pp. 123–129.

- [48] L. J. Muir, I. Richardson, and S. Leaper, "Gaze tracking and its application to video coding for sign language." in *Proceedings of Picture Coding Symposium 2003*. IEEE, 2003.
- [49] D. Mushfieldt, M. Ghaziasgar, and J. Connan, "Robust facial expression recognition in the presence of rotation and partial occlusion," in *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*. ACM, 2013, pp. 186–193.
- [50] 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.
- [51] N. Naidoo and J. Connan, "Gesture recognition using feature vectors," in *Proc. South African Telecommunication Networks and Applications Conference (SATNAC 2009)*, 2009.
- [52] J. Nasiri, S. Khanchi, H. R. Pourreza *et al.*, "Eye detection algorithm on facial color images," in *Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on*. IEEE, 2008, pp. 344–349.
- [53] W. Nel, "An integrated sign language recognition system," Master's thesis, University of the Western Cape, Computer Science, 2014.
- [54] W. Nel, M. Ghaziasgar, and J. Connan, "An integrated sign language recognition system," in *South African Telecommunication Networks and Applications Conference*, 2013, pp. 179–185.
- [55] M. Pantic and L. J. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.
- [56] M. Pardàs and A. Bonafonte, "Facial animation parameters extraction and expression recognition using hidden markov models," *Signal Processing: Image Communication*, vol. 17, no. 9, pp. 675–688, 2002.
- [57] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1449–1456.
- [58] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification." in *nips*, vol. 12, 1999, pp. 547–553.
- [59] 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.

- [60] J. A. Russell, "Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies." *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.
- [61] K. R. Scherer and P. Ekman, *Handbook of methods in nonverbal behavior research*. Cambridge University Press Cambridge, 1982, vol. 2.
- [62] R. Schweiger, P. Bayerl, and H. Neumann, "Neural architecture for temporal emotion classification," in *Affective Dialogue Systems*. Springer, 2004, pp. 49–52.
- [63] M. Sheikh, "Robust recognition of facial expressions on noise degraded facial images," Master's thesis, University of the Western Cape, 2011.
- [64] Y.-l. 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.
- [65] A. Tzotsos and D. Argialas, "Support vector machine classification for object-based image analysis," in *Object-Based Image Analysis*. Springer, 2008, pp. 663–677.
- [66] H. Vadapalli, H. O. Nyongesa, and C. W. Omlin, "Facial action unit recognition using recurrent neural networks." in *IPCV*, 2009, pp. 357–361.
- [67] H. 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 the Western Cape, 2011.
- [68] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society International Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.
- [69] J. Vom Brocke and C. Buddendick, "Reusable conceptual models–requirements based on the design science research paradigm," in *Proceedings of the 1st International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006), Claremont*. Citeseer, 2006, pp. 576–604.
- [70] J. Whitehill, "Automatic real-time facial expression recognition for signed language translation," Master's thesis, University of the Western Cape, Computer Science, 2006.
- [71] J. Whitehill and C. W. Omlin, "Haar features for face recognition," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, p. 5.



- [72] J. Wojdel, A. Wojdel, and L. Rothkrantz, "Analysis of facial expressions based on silhouettes," in *Fifth Annual Conference of ASCI, ASCI, Delft*, 1999.
- [73] Y.-T. Wu, T. Kanade, J. Cohn, and C.-C. Li, "Optical flow estimation using wavelet motion model," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 992–998.
- [74] Y. Yacoob and L. Davis, "Computing spatio-temporal representations of human faces," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference On*. IEEE, 1994, pp. 70–75.

