Unconstrained Facial Expression Recognition in Still Images and Video Sequences using Random Forest Classifiers

Mostafa Abd El Meguid

Master of Engineering

Department of Electrical and Computer Engineering

McGill University

Montreal, Quebec

A Thesis submitted to the Faculty of Graduate Studies and Research in partial fulfilment of the requirements for the degree of Master of Engineering

March 2012

© Mostafa Abd El Meguid
Abstract

The aim of this project is to construct and implement a comprehensive facial expression detection and classification framework through the use of a proprietary face detector (PittPatt) and a novel classifier consisting of a set of Random Forests paired with either support vector machine or k-nearest neighbour labellers. The system should perform at real-time rates under unconstrained image conditions, with no intermediate human intervention. The still-image Binghamton University 3D Facial Expression database was used for training purposes, while a number of other expression-labelled video databases were used for testing. Quantitative evidence for qualitative and intuitive facial expression recognition constitutes the main theoretical contribution to the field.
L'objectif de ce projet est de construire et mettre en œuvre un cadre complète de détection de l'expression du visage par l'utilisation d'un détecteur de visage exclusif (PittPatt) et un nouveau classificateur composé d'un ensemble de 'Random Forests' a accompagné d'un étiqueteur 'support vector machine' ou 'k-nearest neighbour'. Le système doit effectuer au temps réel, dans des conditions sans contrainte, sans aucune intervention humaine intermédiaires. La base de données d'images fixes ‘Binghamton University 3D Facial Expressions’ était utilisé à des fins de formation. Un nombre de bases de données d'expression d'images fixes et de vidéo ont été utilisés pour l'évaluation. Des données quantitatives pour l'analyse qualitative, et parfois intuitive, les sujets liés à l'expression faciale constituaient la contribution principale et théorique sur le terrain.
Acknowledgements

I would like to acknowledge the invaluable guidance provided by my supervisor Professor Martin Levine throughout the formulation, implementation and discussion of this thesis. I would also like to acknowledge the education, and work ethic I gained during my years at McGill University from my numerous instructors and peers.

I would like to thank the members of the Visual Surveillance Group at the Centre for Intelligent Machines, particularly Mohannad Elhamod, for their continuous insight and constructive criticism.

Finally, I would most like to thank my father, my mother, my brother, my extended family and the friends I consider such for their endless presence and support in all its forms, without which I would not be half the person I am.
# Table of Contents

Abstract .................................................................................................................... ii
Abrégé ................................................................................................................... iii
Acknowledgements ............................................................................................................ iv
List of Tables ................................................................................................................ vii
List of Figures ................................................................................................................... ix

I Introduction.................................................................................................. 1
   I-1 Overview ..................................................................................................... 1
   I-2 Thesis Contributions ................................................................................... 5
   I-3 Chapter Contents ........................................................................................ 8

II A Brief History of Emotion and Facial Expressions .................................... 9

III Literature Review ...................................................................................... 11

IV Existing Facial Expression and Emotion Classification Models ................ 21
   IV-1 Pleasure-Arousal-Dominance (PAD) [77] ................................................. 22
   IV-2 Ortony, Clore and Collins (OCC) Model of Emotion [87] .................. 23
   IV-3 Applicability to this Project ........................................................................ 24

V Databases ................................................................................................. 26
   V-1 Binghamton University 3D Facial Expression (BU-3DFE) Database [11] 27
   V-2 Karolinska Directed Emotional Faces (KDEF) Database [94] ............. 27
   V-3 Radboud Faces Database (RaFD) [95] .................................................... 28
   V-4 Japanese Female Facial Expression (JAFFE) Database [96] .................. 28
   V-5 Psychological Image Collection at Stirling – Pain (PICS - pain) [97] 29
   V-6 Static Facial Expressions In the Wild (SFEW) Database [98] ............... 29
   V-7 Binghamton University 3D Dynamic Facial Expression Database [90] 29
   V-8 Facial Expressions and Emotion Database (FEED) [99] ....................... 30
   V-9 Acted Facial Expressions In The Wild (AFEW) Database [100] .......... 30
   V-10 CAS Pose, Expression, Accessory & Lighting (PEAL) Database [101] 31
   V-11 CMU Multi-Pose, Illumination & Expression (Multi-PIE) Database [102] 31
   V-12 MMI Facial Expression Database [103], [104] ........................................ 32
   V-13 Cohn-Kanade AU-Coded Facial Expression Database 2 (CK+) [44] .... 33
   V-14 Summary ................................................................................................... 34

VI Face Detection .......................................................................................... 35
   VI-1 Topics of Concern in Face Detection....................................................... 35
   VI-2 Requirements of Unconstrained Face Detection ..................................... 37
   VI-3 Approaches to Unconstrained Face Detection ....................................... 38
   VI-4 Pittsburgh Pattern Recognition (PittPatt) [132] ...................................... 39

VII Methodology and Implementation ............................................................. 42
<table>
<thead>
<tr>
<th>VII-1</th>
<th>Training Database Preparation</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>VII-2</td>
<td>Feature Analysis and Dimensionality Reduction</td>
<td>51</td>
</tr>
<tr>
<td>VII-3</td>
<td>Training the Classifier</td>
<td>55</td>
</tr>
<tr>
<td>VII-4</td>
<td>Temporal Integration</td>
<td>63</td>
</tr>
<tr>
<td>VII-5</td>
<td>Testing Strategy</td>
<td>65</td>
</tr>
<tr>
<td>VIII</td>
<td>Results</td>
<td>66</td>
</tr>
<tr>
<td>IX</td>
<td>Discussion</td>
<td>67</td>
</tr>
<tr>
<td>IX-1</td>
<td>Expressions and Expression Intensities</td>
<td>67</td>
</tr>
<tr>
<td>IX-2</td>
<td>Still Image Analysis</td>
<td>74</td>
</tr>
<tr>
<td>IX-3</td>
<td>Video Analysis</td>
<td>77</td>
</tr>
<tr>
<td>IX-4</td>
<td>Temporal Integration Scheme Interface</td>
<td>80</td>
</tr>
<tr>
<td>IX-5</td>
<td>Computer Timing</td>
<td>82</td>
</tr>
<tr>
<td>X</td>
<td>Conclusion</td>
<td>83</td>
</tr>
<tr>
<td>Appendix I</td>
<td>References</td>
<td>85</td>
</tr>
<tr>
<td>Appendix II</td>
<td>Methodology and Implementation Flowcharts</td>
<td>95</td>
</tr>
<tr>
<td>Appendix III</td>
<td>Sample Feature Images</td>
<td>99</td>
</tr>
<tr>
<td>Appendix IV</td>
<td>Mean Feature Images</td>
<td>101</td>
</tr>
<tr>
<td>Appendix V</td>
<td>Sobel kNN-Labelling Classifier Accuracies</td>
<td>104</td>
</tr>
<tr>
<td>Appendix VI</td>
<td>Discrete Laplace SVM-Labelling Classifier Accuracies</td>
<td>107</td>
</tr>
<tr>
<td>Appendix VII</td>
<td>Local Binary Pattern SVM-Labelling Classifier Accuracies</td>
<td>109</td>
</tr>
<tr>
<td>Appendix VIII</td>
<td>Still Image Classification Results</td>
<td>111</td>
</tr>
<tr>
<td>Appendix IX</td>
<td>Video Classification Results</td>
<td>121</td>
</tr>
<tr>
<td>Appendix X</td>
<td>Numerical Confusion Matrices for BU-3DFE</td>
<td>127</td>
</tr>
<tr>
<td>Appendix XI</td>
<td>Random Forests</td>
<td>128</td>
</tr>
</tbody>
</table>
List of Tables

Table 1: Action Units in the Facial Action Coding System [43] (Table adapted from [51]) .......... 18
Table 2: Summary of databases as used in this project................................................................. 34
Table 3: Numerical breakdown of RIS contents ........................................................................ 47
Table 4: Potential landmarks extracted by PittPatt in each of 3 yaw ranges (adapted from [1]) . 47
Table 5: Numerical breakdown of the VIS contents according to expression and intensity........ 49
Table 6: Cropping parameters for the 12 yaw bins. The dashed black box denotes the face region detected by PittPatt with height $H$ and width $W$ (adapted from [1]). LEX = Left Eye $x$ coordinate, LEY = Left eye $y$ coordinate, RE = Right Eye, NBR = Nose Bridge, NBA = Nose Base .......... 50
Table 7: Face resize dimensions for the 12 yaw bins (all measurements are in pixels).............. 51
Table 8: Feature Analysis Operators.......................................................................................... 52
Table 9: Mean feature Image and first 5 eigenfaces for each feature in yaw bin 6 ...................... 54
Table 10: Parameters used for RF construction......................................................................... 56
Table 11: Central bins (5, 6, 7 and 8) mean classification accuracies for Sobel feature Classifiers with SVM labelling (all values are %) ......................................................................................... 62
Table 12: Central bins (5, 6, 7 and 8) mean classification accuracies for Sobel feature Classifiers with k-NN labelling (all values are %) ............................................................................................ 62
Table 13: Optimal classifier structure for SVM based labelling.................................................. 63
Table 14: Best central bins (5, 6, 7, and 8) accuracy obtained for each of the still image test databases along with the classifier parameters giving rise to this accuracy................................. 75
Table 15: Classification accuracy achieved by other methods ................................................... 75
Table 16: Classification accuracy of the different versions of the framework, compiled from [148] ..................................................................................................................................................... 76
Table 17: Video databases evaluated in this thesis .................................................................... 77
Table 18: Classification accuracy achieved by other methods (all values are in %).................... 78
Table 19: Results of classifying the BU-4DFE by the 2D method in [149]................................. 79
Table 20: Results of classifying the BU-4DFE by the 3D method in [150]................................. 79
Table 21: Characteristics of the computer system used for training, testing and storage of the classifier ......................................................................................................................................... 82
Table 22: Sample image from the CRIS (left) and from the original BU-3DFE database (right) with their respective S, DL and LBP feature images. The full sized texture images from the BU-3DFE database were included here simply for visual purposes................................................................. 100
Table 23: Mean Sobel feature images for the 7 expressions in the 12 yaw bins ....................... 101
Table 24: Mean Laplace feature images for the 7 expressions in the 12 yaw bins .................... 102
Table 25: Mean Local Binary Pattern feature images for the 7 expressions in the 12 yaw bins 103
Table 26: Central bins (5, 6, 7 and 8) mean classification accuracies for Discrete Laplace feature
Classifiers with SVM labelling (all values are %) ................................................................. 108
Table 27: Central bins (5, 6, 7 and 8) mean classification accuracies for Local Binary Pattern
feature classifiers with SVM labelling (all values are %) ...................................................... 110
Table 28: Numerical values related to the Sobel confusion matrix in Figure 24 (all values are %)
................................................................................................................................................ 127
Table 29: Numerical values related to the discrete Laplace confusion matrix in Figure 24 (all
values are %) ...................................................................................................................... 127
Table 30: Numerical values related to the local binary pattern confusion matrix in Figure 24 (all
values are %) ...................................................................................................................... 127
List of Figures

Figure 1: Structure of a typical facial expression recognition system ................................................. 4
Figure 2: Noisy Mona Lisa images with the authors’ suggested underlying emotion for each case [24] ................................................................................................................................................. 12
Figure 3: ‘Constrained Local Model’ fitting scheme; an exhaustive local search for feature locations is followed by a stochastic optimisation strategy to maximise the responses (or likelihood) of one of the Point Distribution Models on the right [39] ................................................................................................. 16
Figure 4: Point Distribution Model (left), Superimposed Geometric Measurements for the Active Shape Model (right) [53] ................................................................................................................................................. 18
Figure 6: Yaw variation for subject ‘F0003’ displaying a neutral expression .................................... 44
Figure 7: Subject ‘F0001’ displaying expressions representing ‘anger, disgust, fear, happiness, sadness and surprise’ with maximal intensity, as well as the neutral expression ............................. 44
Figure 8: Subject ‘M0038’ displaying expressions representing ‘anger, disgust, fear, happiness, sadness and surprise’ with maximal intensity, as well as the neutral expression ....................... 45
Figure 9: ‘F0001’ displaying expressions representing ‘anger, disgust, fear, happiness, sadness and surprise’ at increasing intensities from left to right ........................................................................ 46
Figure 11: Histogram images per yaw bin before (red) and after (blue) horizontal flipping in the VIS ................................................................................................................................................. 48
Figure 10: Yaw bin limits and sample images in the VIS ....................................................................... 48
Figure 12: Sample cropping process for a face in the VIS. Original image (left), cropped image (centre), final cropped and resized image (right) ........................................................................... 51
Figure 13: Per pixel grey scale intensity reconstruction error for varying number of components 53
Figure 14: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 100 trees per forest under the three different labelling schemes using Sobel features ........ 59
Figure 15: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 100 trees per forest under the three different labelling schemes using Sobel features .... 59
Figure 16: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 100 trees per forest using Sobel features .............................................................................. 59
Figure 17: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 200 trees per forest under the three different labelling schemes using Sobel features .... 60
Figure 18: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 200 trees per forest under the three different labelling schemes using Sobel features .... 60
Figure 19: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 200 trees per forest using Sobel features .............................................................................. 60
Figure 20: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 300 trees per forest under the three different labelling schemes using Sobel features ................................. 61
Figure 21: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 300 trees per forest under the three different labelling schemes using Sobel features ........................................ 61
Figure 22: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 300 trees per forest using Sobel features ..................................................................................... 61
Figure 23: Temporal Integration Scheme Visualisation ................................................................... 64
Figure 24: Sobel (left), discrete Laplace (centre) and local binary pattern (right) expression confusion matrix. Note that darker shades indicate higher counts.......................................................... 67
Figure 25: Correlation coefficients for the 7 expressions. Note: the correlation coefficient sum (Σ) does not include self-correlation ........................................................................... 68
Figure 26: Confusion matrix for the RaFD. The values come from human evaluation of images [95] ........................................................................................................................................ 70
Figure 27: Confusion matrix for type 1 RF collections (left) and type 2 RF collections (right) with 300 trees/RF using Sobel features and type 1 labelling scheme .............................................. 71
Figure 28: Binary correlation coefficients for the 25 intensities calculated from the confusion matrix on the right in Figure 27. Black represents positive correlation, white represents negative correlation. .................................................................................................................. 72
Figure 29: Lowest intensity face image from BU-3DFE database for [anger, disgust, neutral, and sadness] and [fear and happiness] displayed by subject ‘M0014’ ......................................................... 72
Figure 30: Auto-confusion matrices for anger, disgust and happiness taken from the matrix on the right in Figure 27 .................................................................................................................. 73
Figure 31: Sample interface of the temporal integration scheme ........................................................................................................ 81
Figure 32: Worst case scenario timing diagram ................................................................................. 82
Figure 33: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 100 trees per forest under the two different labelling schemes using Sobel features ........... 104
Figure 34: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 100 trees per forest under the two different labelling schemes using Sobel features ........... 104
Figure 35: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 100 trees per forest using Sobel features .................................................................................. 104
Figure 36: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 200 trees per forest under the two different labelling schemes using Sobel features ........... 105
Figure 37: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 200 trees per forest under the two different labelling schemes using Sobel features ........... 105
Figure 38: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 200 trees per forest using Sobel features .................................................................................. 105
Figure 39: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 300 trees per forest under the two different labelling schemes using Sobel features

Figure 40: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 300 trees per forest under the two different labelling schemes using Sobel features

Figure 41: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 300 trees per forest using Sobel features

Figure 42: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 200 trees per forest under the two different labelling schemes using Discrete Laplace features

Figure 43: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 200 trees per forest under the two different labelling schemes using Discrete Laplace features

Figure 44: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 200 trees per forest using Discrete Laplace features

Figure 45: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 300 trees per forest under the two different labelling schemes using Discrete Laplace features

Figure 46: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 300 trees per forest under the two different labelling schemes using Discrete Laplace features

Figure 47: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 300 trees per forest using Discrete Laplace features

Figure 48: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 200 trees per forest under the two different labelling schemes using Local Binary Pattern features

Figure 49: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 200 trees per forest under the two different labelling schemes using Local Binary Pattern features

Figure 50: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 200 trees per forest using Local Binary Pattern features

Figure 51: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 300 trees per forest under the two different labelling schemes using Local Binary Pattern features

Figure 52: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 300 trees per forest under the two different labelling schemes using Local Binary Pattern features
Figure 53: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 300 trees per forest using Local Binary Pattern features .............................................................. 110

Figure 54: Number of face images found per yaw bin according to true expression label for KDEF ..................................................................................................................................................... 111

Figure 55: Classification accuracy of KDEF using the optimal Sobel based classifier from Table 13 ................................................................................................................................................. 111

Figure 56: Classification accuracy of KDEF using the optimal Discrete Laplace based classifier from Table 13 ............................................................................................................................................. 112

Figure 57: Classification accuracy of KDEF using the optimal LBP based classifier from Table 13 ..................................................................................................................................................... 112

Figure 58: Number of face images found per yaw bin according to true expression label for RaFD ..................................................................................................................................................... 113

Figure 59 Classification accuracy of RaFD using the optimal Sobel based classifier from Table 13 ..................................................................................................................................................... 113

Figure 60: Classification accuracy of RaFD using the optimal Discrete Laplace based classifier from Table 13 ............................................................................................................................................. 114

Figure 61: Classification accuracy of RaFD using the optimal LBP based classifier from Table 13 ..................................................................................................................................................... 114

Figure 62: Number of face images found per yaw bin according to true expression label for JAFFE .......................................................................................................................................... 115

Figure 63: Classification accuracy of JAFFE using the optimal Sobel based classifier from Table 13 ................................................................................................................................................. 115

Figure 64: Classification accuracy of JAFFE using the optimal Discrete Laplace based classifier from Table 13 ............................................................................................................................................. 116

Figure 65: Classification accuracy of JAFFE using the optimal LBP based classifier from Table 13 ..................................................................................................................................................... 116

Figure 66: Number of face images found per yaw bin according to true expression label for PICS-pain .............................................................................................................................................. 117

Figure 67 Classification accuracy of PICS-pain using the optimal Sobel based classifier from Table 13 ..................................................................................................................................................... 117

Figure 68: Classification accuracy of PICS-pain using the optimal Discrete Laplace based classifier from Table 13 ............................................................................................................................................. 118

Figure 69: Classification accuracy of PICS-pain using the optimal LBP based classifier from Table 13 ..................................................................................................................................................... 118

Figure 70: Number of face images found per yaw bin according to true expression label for SFEW ..................................................................................................................................................... 119
Figure 71: Classification accuracy of SFEW using the optimal Sobel based classifier from Table 13 ........................................................................................................................................................................... 119
Figure 72: Classification accuracy of SFEW using the optimal Discrete Laplace based classifier from Table 13 ........................................................................................................................................................................... 120
Figure 73: Classification accuracy of SFEW using the optimal LBP based classifier from Table 13 ........................................................................................................................................................................... 120
Figure 74: Number of video sequences tested according to true expression label for BU-4DFE ........................................................................................................................................................................................................................... 121
Figure 75: Classification accuracy of BU-4DFE using the optimal Sobel based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 121
Figure 76: Classification accuracy of BU-4DFE using the optimal Discrete Laplace based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 122
Figure 77: Classification accuracy of BU-4DFE using the optimal LBP based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 122
Figure 78: Number of video sequences tested according to true expression label for FEED ........................................................................................................................................................................................................................... 123
Figure 79: Classification accuracy of FEED using the optimal Sobel based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 123
Figure 80: Classification accuracy of FEED using the optimal Discrete Laplace based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 124
Figure 81: Classification accuracy of FEED using the optimal LBP based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 124
Figure 82: Number of video sequences tested according to true expression label for AFEW ........................................................................................................................................................................................................................... 125
Figure 83: Classification accuracy of AFEW using the optimal Sobel based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 125
Figure 84: Classification accuracy of AFEW using the optimal Discrete Laplace based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 126
Figure 85: Classification accuracy of AFEW using the optimal LBP based classifier (with labelling schemes 2 and 3) from Table 13 ........................................................................................................................................................................................................................... 126
I Introduction

I-1 Overview

The goal of this thesis was to design and implement a functional framework for the classification of facial expressions present on the faces of individuals in still images and videos. The major premise is that externally expressed facial expressions are a major window on internal emotions. We also assumed that the framework was required to operate with neither human intervention nor constraints with respect to demographics, physical pose, or background environment. Facial expressions join a plethora of human behaviours applicable to fields such as marketing, security, clinical research, and human computer interaction.

The theoretical roots of this project lie in the work of Paul Ekman and his various collaborators over the years. Their six ‘universal expression’ model [2] comprised of anger, disgust, fear, happiness, surprise and sadness embodies the solution space of the classification problem. These ‘universal expressions’ are assumed to communicate the same message across the human race, void of cultural, geographical or temporal bias. In turn, Ekman was influenced by earlier psychological and philosophical studies, most notably by René Descartes, Charles Darwin, William James, and John Dewey. The opinions of these pillars in their fields varied with respect to the causes behind emotions, with the biggest two questions being what produces emotions and how are they expressed?

In its purest form, the link between the emotion felt by an individual and the facial expression they display is both reflexive and reflective. However, expressions are generally accepted to be the primary observable response. Thus, for computer vision purposes, the shortest route to investigating human emotion cuts directly through the field of facial expression research. Naturally, human emotions are an area of great interest for many practical applications, rendering facial expression research a popular topic.

Unfortunately, much like any other human behavioural metric, facial expressions suffer from the diversity of mankind. The minor nuances and variations in the displayed expressions may be comprehended by humans; however, for a computer vision system each variation represents an exception to the rule. Humans use context to differentiate expressions that may independently convey more than one message. Herein lies the limiting factor to the widespread application of facial expression detectors; the luxury of context is not as easily encoded into an artificial machine.
The topic of *unconstrained* facial expression recognition was chosen due to the relative lack of attention it has received in the research community. Publications detailing a fully automated real-time classifier are virtually non-existent. Tian, Kanade and Cohn pointed out in their thorough 2011 review of the state of facial expression research that 'In the face expression literature, use of multiple perspectives is rare; and relatively less attention has been focused on the problem of pose invariance' [3]. Most prominently, minor experimentation somewhat related to unconstrained facial expression recognition was included in [4]. Here, the authors developed a 3D facial model that was fitted to 2D facial images in an attempt to track eye regions. According to the authors, this experiment served as a precursory probe into full blown facial expression classification. Similarly, commercial facial expression detectors such as [5] focus on frontal, camera-facing poses due to their involvement in product evaluation by consumers in a marketing setting.

Aside from performing at real-time rates and classifying expressions under a wide range of head poses, an appropriate framework should not be hindered by confounding characteristics such as age, gender, race or skin colour. In addition, a *confidence measure* of the detected expression, along with an *expression intensity* rating, would give the framework output more substance and meaning. Expression intensity and confidence are certainly not novel concepts in the field, with studies such as [6] and [7] attempting to ascertain values for these parameters, albeit in heavily constrained environments. For the purposes of this thesis, these extensions to a basic detection framework were treated as essential requirements. In particular, expression intensities were made intrinsic to the overall framework, from the training stage all the way through to final expression labelling.

Identifying any specific type of system failure that generalises to all current facial expression classifiers would trivialise the problem. Difficulties are encountered at every step, and in most cases, they propagate through to the final result. One seemingly trivial example of a far reaching problem that is commonly encountered involves the model chosen to represent or visualise the results. The model needs to incorporate multiple expressions and expression intensities. It can take the shape of coordinate axes, a probabilistic distribution or any other mathematical space. The classifier used to evaluate an input facial image should produce a result that can be translated into a structure that the model can represent. Thus, streamlining the data flow from an input image to a singular point or cluster in the expression model is one of the largest challenges.

In addition, the training images used to develop the system by training the classifier need to accurately and comprehensively represent the mathematical model while maintaining generalisability for future use. This requires that the model be kept in mind when selecting the
training set, highlighting the importance and close relationship between the different subcomponents of the overall classifier.

Facial expression classification could be considered as being related to face detection, since clearly the former will influence ability to achieve the latter. In terms of accuracy of detection, pose invariance and demographic accommodation, near perfect (>99% accuracy) face detection has been achieved even under difficult conditions (real-time performance, low resolution of images, small face dimensions, varying demographics and varying pose) as reported in [8]. Meanwhile, the highest classification rates to date for the six Ekman expressions in real-time is 91.5% in [9]. However, the authors of [9] reported these accuracy rates for strictly frontal images, which were unfortunately selected from the same image set that was used to train their system. This renders their results quite limited, as will be expanded upon later in the thesis. The highest reported detection rates for a facial expression classifier tested on image sets entirely uninvolved in the training stage drops to around 70% as reported in [3]. It should be noted that none of these reported methods is a completely automated, multi-pose and real-time facial expression recognition system.

The general structure of a typical facial expression recognition system, shown in Figure 1, begins with the detection and acquisition of a face from an image or video. At this stage, a number of characteristics of the detected face are extracted. These can include head pose, face size, and facial landmark locations. Landmarks can be related to any facial features such as the eyes, nose and mouth. The extracted characteristics can then be used to describe and analyze the face. Image normalisation may precede the classification stage to match the input image dimensions to training samples. More often than not, the face detection stage is left to a proven or commercial framework, allowing researchers to focus on the core problem of expression classification. Examples of face detectors and their subcomponents used in expression classification frameworks are provided in Chapter III and Chapter VI.

Our approach to face detection employs a commercial detector offered by PittPatt [1] obtained under a full academic licence. PittPatt provides metadata for each detected face comprising of pose, size and facial landmarks. Once a face has been detected, it is normalised with respect to brightness, contrast, size and pose. Subsequently, feature analysis and dimensionality reduction of the face attempts to map it into a space where expressions are more easily identified.

1 Since obtaining this software, PittPatt has been bought out by Google.
Analysis follows image processing and generally is achieved by one of two approaches. *Appearance-based* analysis methods use feature analysis in the form of image filters to map the face into a new space, one where expressions may be more easily identified using machine learning tools. *Model-based* analysis methods use the location, shape and intermediate distances of the detected landmarks to produce a semantic descriptor. The result of the analysis stage can then be used by a classifier to determine the expression label. If the input medium is a video or sequence of images, information should be propagated from one frame to the next.

The classifier used in this project consists of two stages: *evaluation* and *labelling*. The evaluator was chosen to be a collection of Random Forest (RF) classifiers and the labeller, a support vector machine (SVM). The RF collection as presented here is an *entirely novel* structure that benefits from the advantages of individual RFs that Leo Breiman outlined in [10], while breaking down a multi-label classification problem into a number of more manageable binary problems. The labeller takes the output of the RF collections and extracts a single, most likely, expression from all of the available possibilities.

One of the biggest challenges in the actual implementation of our framework was the acquisition of a comprehensive database used to build the classifier. As discussed above, the selection of the database is highly related to the potential success of the overall framework. The Binghamton University 3D Facial Expression database (BU-3DFE) [11] contains three dimensional, textured face models displaying a wide array of expressions. This very flexible database was used to produce a 2D training image set covering the full repertoire of expressions, expression intensities and head poses to be classified. However, BU-3DFE has one major drawback. It was not obtained under unconstrained conditions.

---

2 A k-nearest neighbours (k-NN) labeller was also tested.
One of the most important discussions in the literature involving facial expression databases arises from the psychological theories behind the origin of emotions and expressions. Most databases available for research purposes were captured in laboratory environments where the subjects were directed to display a certain expression. These non-spontaneous or acted databases differ from their spontaneous counterparts where facial expressions were caused by actual stimuli. In most cases the stimulus was a video aimed at stirring certain emotions in the subjects. Arguments against using a non-spontaneous database for unconstrained, spontaneous facial expressions call upon the possible falseness of the non-spontaneous sets, particularly the loss of the involuntary portions of the expressions. Unfortunately, as expanded upon in Chapter V, no comprehensive spontaneous image set for training a classifier is currently available. Chapter V provides a thorough breakdown of facial expression databases, as well as issues related to their use.

I-2 Thesis Contributions

A number of novel methods were introduced in this project with no record of their previous appearance in the literature:

Image Processing:
   As mentioned above, this stage prepares the detected face for expression analysis. In most cases, the analysis and subsequent classifier are designed to handle inputs of particular dimensions and orientations.

   Alignment and orientation of images is of particular importance for appearance-based approaches where spatial coherence is of the utmost importance. Polynomial warping used in facial expression implementations, extensively described in [12] and [13], provides this spatial coherence. However, the computational cost of evaluating every new image’s polynomial mapping conflicts with the real-time requirements of our system.

   In our framework, alignment of images did not involve any nonlinear warping; instead the detected coordinates of the facial landmarks were used to define a simple rectangular crop window. Resizing of detected faces was controlled by the dimensions of training samples sharing the same pose. Thus, simple arithmetic operations requiring negligible processing time were involved in the alignment step.
The positive results, evident in the mean images presented in Appendix III, indicate that the strength of the analysis and classification stages can leverage the accuracy lost due to using a simplified alignment procedure.

Classifier:

In almost all facial expression classification frameworks, the classifier takes an abstraction of the input image in the form of a feature transformation or landmark coordinates and distances. Thus it is a combination of the abstraction and the classifier that define the ability of the framework to properly identify a large number of expressions. A detailed description of abstractions and classifiers used in the literature is available in Chapter III.

Our abstraction consists of a dimension-reduced version obtained by image filtering, using common tools available in the image processing and computer vision literature. The resulting image abstraction is a one-dimensional vector which is input to a decision tree.

The classifier as described earlier can be broken down into two parts, an evaluator and a labeller. The evaluator, consisting of a collection of RFs, is unique in its structure. The breakdown of the problem of classifying an expression as one of a large number of possibilities into a set of binary problems is entirely novel. Each RF was trained to classify an input as either displaying a certain expression or not. The output of each forest was a confidence value in the respective expression. Combining these confidence values using the aforementioned SVM labeller is also novel.

Sequential Tracking:

Videos encode significant information that is otherwise missing from still images. Temporal progressions may help identify facial expressions more than individual still images. Extraction of this temporal information can take two forms, the first of which requires that the analysis tools and classifier can accommodate sequential data. The most popular methods for this type of classification make use of optical flow as discussed in [14] and [15].

The second approach to incorporating video information leaves the task to a data fusion scheme following the classification of the individual frames. This was the approach we chose.

3 The six Ekman expressions, each at four intensities, plus a neutral expression.
Classification of facial expressions in videos by a system trained on still images such as ours requires a method for combining information from frame to frame to produce a stable, reliable label. This method, called a temporal integration scheme here, uses histograms of the single frame classification results to produce smooth transitions in the final single expression label and quell anomalous expression classifications.

In addition, several experimental outcomes in Chapter IX based on testing the full framework have not previously appeared in the literature:

Inter-expression Correlation

Although many works in the literature have adopted Ekman’s theories, none have quantified the similarities between the six ‘universal’ expressions in the form explicated later in this thesis. Extracting these correlation values identifies the difficulty of classifying each expression with respect to its counterparts.

Intra-expression (or Expression Intensity) Correlation

We provide new evidence for the existence of strong relationships between intensities of the same expression, a topic which is nonexistent in the literature. In addition, evidence for the proximity of low intensities of different expressions is provided.

In summary, the contributions discussed in this thesis introduce new methods for approaching several common subtasks associated with the expression classification problem. We also present some new experimental results for unconstrained facial expression recognition.

In addition to the above stated contribution, we also provide evidence supporting and opposing certain results and theories appearing in the literature. These observations, along with the correlation findings, give support to the methodology employed in this thesis, in particular, the two pivotal stages in the classification process. One is the pivotal analysis stage, which represents the connective link between theory and the actual practical implementation. We note that further investigation into methods for input image manipulation is perhaps integral to obtaining better expression classification outcomes.

The structure of the classifier is the second area where further investigation is required. The solution space to the facial expression problem is quite large, and the complexity is exacerbated once sequential information, such as video, is incorporated. Thus, ensuring that a classifier used
to evaluate the images can handle this large problem should be a priority for researchers in this field.

I-3 Chapter Contents

This thesis takes the reader through the history of facial expression research in Chapters II and III. The focus then shifts to the subcomponents of a facial expression classifier, where the underlying models for expression classification are summarised in Chapter IV. Chapter V contains a critique of the databases available for facial expression research. An overall breakdown of face detection and how it relates to expression recognition is provided in Chapter VI.

A walkthrough of the overall methodology, implementation and reasoning behind certain design choices is contained in Chapter VII. The results of system testing are provided in Chapter VIII. A discussion of these results follows in Chapter IX, capped by a conclusion of the overall outcome in Chapter X.

The appendices include a reference list, average expression images under various image filters, classifier accuracy plots, flow charts detailing each stage of the framework and finally an extended review of Random Forests.
II  A Brief History of Emotion and Facial Expressions

Emotion in humans has been an active area of modern research from as far back as the late nineteenth century when Charles Darwin deemed emotions and facial expressions to be a communication tool used by humans and animals alike. He especially noted that the structure of the nervous system has much to do with involuntary expressions and actions [16].

Darwin’s latter point regarding physiological relevance was bolstered by the seminal work of William James in ‘The Emotions’ chapter of his ‘Principles of Psychology’ textbook [17]. In the midst of producing the manuscript for this book [18], James pre-emptively published an article in the ‘Mind’ journal [19] where he purveyed some of his strongest views, including the opinion that emotions are the state of the conscious mind following an event to which the body had reacted. The importance of the ‘consciousness of the mind’ comes into play when true reactionary facial expressions are considered, which may precede the final emotional state. Thus, both Darwin and James agreed that internal emotions are not the purest form of the body’s reaction to a stimulus; rather it is the way in which these emotions are expressed, be it facially, bodily or otherwise observable that reflects the true underlying affective state. Although this project does not necessarily follow the ordering proposed by these two giants of humanistic science, the significance of their findings comes into play in the discussion of the databases (Chapter V) used to construct and test the recognition and classification system.

In 1895, John Dewey attempted to close the gap between the works mentioned above in a journal article titled the ‘The Theory of Emotion’ [20], suggesting the necessary minor changes to the various statements made by each author to fall in accordance with their counterparts work. We see Dewey’s largest contribution lying in his ‘third head’ (p. 560), where he highlighted the ‘teleological’ nature of expressions, concurring with James’ work, while requiring Darwin’s general ‘principle of serviceable associated habits’ to not simply be a reflexive property such as the case of elation automatically bringing about laughter or smiles, but also goal-oriented. Dewey’s caveat may seem trivial; however, his insistence on pairing expression with ‘goal’ is greatly supported by Paul Ekman and Wallace Friesen’s work which in turn has spurred a plethora of both computer vision and recent psychological research.

Expression under Dewey is partially defined as a communication tool, attempting to convey (or conceal) the emotion felt by the subject. Successful communication in any scenario requires that source and sink follow a common protocol, part of which is the language of the medium. Ekman

---

4 Also known as the James-Lange theory after Danish psychologist Carl Lange’s parallel work
and Friesen [21] investigated the domain of nonverbal communication between humans, questioning the intended and actual result of their display. This approach could be seen as an examination of Dewey’s aforementioned addendum and the series of studies referenced in [22] may support a biological predisposition in humans for understanding this type of communication.

One of Ekman and Friesen’s early and well-established findings was the ‘universality’ of certain facial expressions (p. 71) [23]. Seven emotions labelled as the ‘primary affects’ were found to be consistently identifiable through the reflexive facial expressions they bring about. These affective states are ‘happiness, surprise, fear, sadness, anger, disgust and interest’, nowadays commonly reduced to six through the exclusion of ‘interest’.

Ekman and Friesen discussed the ‘origin’ of said six emotions, particularly the Darwinian explanation behind them and the associated facial expressions. They acknowledged the evolutionary relevance of some observations, such as wide open eyes in situations bringing about a fearful or distressed affective state being used to increase the field of vision to better spot predators, or a baby crying when it is unhappy with something to attract the attention of those around. However, they were dissatisfied with a purely survivalist explanation due to its inability to clearly track the development of subtleties such as smiles and laughter used to express happiness, or a furrowed brow to express sadness or anger. Thus, Ekman and Friesen attributed parts of the human facial expression repertoire to a communicative characteristic of the species, one that has managed to cross geographic and cultural boundaries.

Another interesting aspect to the ‘panculturalism’ of facial expressions is the ‘evoking stimuli’. Ekman and Friesen suggest that the same situations in two disjoint cultures could evoke different emotions and expressions in the two sets of individuals. Thus it is of utmost importance to segregate emotion from stimulus when attempting to bridge expression perception across mankind. Scenarios showcasing this point are numerous, with one simply needing to draw on conflicting cultural customs and the subsequent reaction of those around. Admittedly, in today’s globalised community, realisation of cultural differences has led many to accept others’ reactions as norm, particularly if they appear to come from a different background to themselves.
III Literature Review

The important role of the mammalian face in communication could be seen as a subsequence of the intricate muscular structure that allows for the display of a multitude of accurate, measured expressions. In fact, facial expressions are at times more communicative than any other medium an individual may use to convey their inner emotional state to an external agent. Species from across the animal kingdom employ facial expressions as part of their comprehensive body language, and humans are especially of no exception. Whether the observed expressions are a biological ingrain across mankind or a cultural by-product has been a topic of debate for years. Regardless of their true origin, there can be no denying that much could be extracted from the subtle changes in the arrangement of facial organs that may not otherwise be communicated. In the same sense, facial expressions can insert uncertainty into an exchange when conflicting messages are purveyed by one agent to the other(s). An example showcasing the effect of minor changes in facial expression and the perceived underlying emotion comes from the realm of classical Renaissance art. Leonardo da Vinci’s Mona Lisa is one of the most admired paintings in history, drawing millions of visitors every year to the Louvre in Paris. One of the most intriguing features of the legendary portrait is the non-descript facial expression the subject appears to bear. While proponents of each opinion may attempt to reason their choice through historic references and expertise on the art piece’s contemporaries, most agree on the subtlety of the expression depicted in the painting. Kontsevich and Tyler [24] performed a novel experiment on a greyscale version of the face in the Mona Lisa. After adding random Gaussian noise to the face on several occasions, the authors provided their opinion on the resultant expressions as shown in Figure 2. Although the suggested emotions are not obtained through a ‘systematic study’, the variations and subsequent labels are captivating in their similar appearance yet varied perception of emotion.
The issue of expression subtlety will resurface later on during discussion of the way expressions and their implied emotions are mapped into a continuous geometric space. For now, remaining in the realm of observational science, the reader is asked to ignore the labels in Figure 2 and attempt to provide their own caption for each case. Seldom would the chosen words match precisely the ones picked by Kontsevich and Tyler, regardless of their synonymy. Assuming that the reader's labels were picked in English, this semblance would be within a single language, leading to the issue of infiniteness of the word selection used to describe any expression or emotion across cultures. This large vocabulary may serve a purpose in literary fields, adding beauty to writings through variety; however, rarely can these word sets be ordered on a continuous scale that is agreed upon by all members of a single culture or speakers of a language, never mind across different ones.

With facial expressions and their causal emotional states playing prominent roles in the human psyche, it was only natural for psychologists to lead the way in attempting to construct a hard science out of the subject. However, the problems described in the previous paragraph required a solution, one that has not necessarily been reached to date. Attempts to reduce the range of
human emotions and expressions to a numbered few are copious and draw upon a variety of studies. Historically, the discussion of emotions and expressions, facial or otherwise, stemmed from the same branch that influenced early research on almost all biological topics over the past 15 years, evolutionary theory. Naturally, criticism of Darwin’s theory’s relevance spawned a number of new publications, with some containing their author’s opinion on a constrained emotion list. Ortony and Turner’s [25] compilation of the ‘basic emotions’ listed in said publications, along with the added references in [26], demonstrate the lack of any collective agreement amongst academics in the field. One emotion model that has been continuously referenced, and somewhat combines previous seminal work in the field is that developed by Ekman and his various collaborators over the years [21].

Ekman’s model is based on the premise that there exist six emotions, the communication of which through facial expressions can be ‘universally recognised’. The six emotions, ‘happiness, surprise, fear, sadness, anger, and disgust’, were initially picked following interaction with isolated tribes in Papua New Guinea [2]. The claimed ‘universality’ of these emotions allows for their employment as dimensions for scientific research, an integral pillar of this project as shown in Section Error! Reference source not found.. Furthermore, the relationship (and at times, assumed interchangeability) between the internal, invisible emotions and their related, observable facial expressions permits for studies concerned with the latter to be built based on Ekman’s theories.

Aside from the noble goal of better understanding the human mind, the domain of emotion research, inclusive of facial expressions, has much to offer in the materialistic world. A standard introduction precedes most recent publications concerned with these topics, listing the shift towards more multi-faceted human computer interaction (HCI) or man-machine interaction (MMI) systems, security implications in detecting deceptive behaviour (sensationalised in the television program ‘Lie to Me’ [27]), as well as psychological analysis for medical purposes. Dwelling on specific applications for the proposed real-time, unconstrained, facial expression recognition system would serve little purpose beyond rehashing the opening statement of a number of the papers referenced in this report.

At its core, automatic facial expression recognition is a standard applied computer vision problem involving all familiar stages of an intelligent, learn by example system. Characteristics that differentiate one algorithm used for solving a computer vision problem from its counterparts lie in a countable number of areas. The four main components that differentiate solutions to a problem within the broader field of computer vision are the sources of training data, the feature analysis techniques used (if any), the approach used (appearance-based vs. model-based) and the machine learning tool employed. The review of existing facial expression recognition systems
included in this report extends the general list of differentiating aspects to include the emotion model and mapping space used (see Chapter IV).

Setting a clear vocabulary to be used hereafter is essential for understanding the work reported as well as for creating a platform for theory and result comparison within the field. It is not uncommon for authors to use synonyms to describe even the most technical portions of their research, an unfortunate subsequence of the lack of clear, set-in-stone theory. *A project at the World Wide Web consortium to develop an ‘Emotion Markup Language’ has not yet been adopted as a standard* [28]. Throughout this report, when *emotion* is referenced, it refers to the internal *affective state* that an individual experiences. Examples of affective states include Ekman’s six basic emotions to their varying degrees of intensity. Meanwhile, ‘facial expressions’ or at times simply ‘expressions’ are the physical manifestation of these emotions. These definitions may not completely agree with those discussed in Chapter II since the system constructed here assumes that emotions and related facial expressions are connected through a two-way implicative relationship.

The previous works reviewed here are mainly concerned with facial expression detection in *unconstrained* environments. The commentary provided summarises both theoretical and technical approaches selected by the authors. Further insight into the pros and cons of the various choices, including their relevance to our proposed unconstrained facial expressions detection system are available in Chapter VI. In addition, the review focuses on more recent attempts at detection and classification and we invite the reader to visit surveys of earlier methods included in [29] and partially in [30].

An objective, computer vision look at the problem of facial expression detection brings about a quizzical *chicken and egg* conundrum. Does the face upon which the expression lies need to be identified before the expression is found, or can the expression speak of its own nature, as an individual entity within an image? A good extension, of this question is whether computer vision researchers strictly require head and neck localisation before faces are detected, the answer to which is certainly not. However, any prior knowledge of head pose and location, for example, in mug shots, invariably assists the face detection process both in speed and accuracy. Facial expression detection could abide by a similar logical argument, where knowledge of face location minimises the search space and almost ensures the existence of a solution (eliminates the null solution since every face must have an expression).

Prior to face detection, most facial expression detection systems require a database for training purposes. The characteristics of an ideal database vary with regard to the goal of the system;
however, *multi-posed subjects* bearing graded changes in expressions are standard prerequisites within this field. A detailed discussion of available databases is provided in Chapter V. A point of concern with most databases, as mentioned in [31], is the ‘forced’ nature of the expressions displayed in the images. By requiring subjects to act out expressions, involuntary portions of the true expressions are eliminated. Meanwhile, a more psychological argument would involve the lack of a true affective state forcing or accompanying these expressions, rendering the expressions false or deceptive. Unfortunately, to date, no full database containing all-natural facial expressions at various poses has been published.

Face detection has arguably been the darling topic for computer vision researchers, with very high detection rates at real-time speeds in unconstrained environments becoming the standard expectation of any algorithm. A detailed summary of the components required for an acceptable face detector are given in Chapter VI. Although the same terminology may be used amongst detection methods, with similarities existing at the bigger picture level, it is important to acknowledge that subtle variations in the interpretation of these approaches will exist.

A well-guided drift in the facial expression research field has involved outsourcing the search for a face location in an image to an existing face detection framework. Examples of this design choice include Essa’s work [31] where Pentland et al.’s eigenspace methods [32] for face detection were utilised. Littlewort et al. [33] customised and implemented the ever popular Viola-Jones face detector [34]. Matsugu et al. [35] used a modified version of the convolutional neural network developed by Fasel in [36]. These three detectors represent a distinct pre-processing step, the lone result of which is a reduced search space for the expression classifier, in addition to metadata such as the size and pose of the detected face. In [37] and [38], attempts to fit geometric models representing faces to image regions serves two purposes; first and more obviously, successful mapping of said model to an image region denotes the localisation of a face. Secondly, the fitted model’s parameters could serve as the features or descriptors for use in the subsequent classification algorithm.

The methods discussed above, with varying modifications, are implemented in one form or another in most comprehensive facial expression applications. However, the importance of the result obtained from the face detector will vary from system to system. One particular case with heightened importance of the face detector is highlighted at the end of the previous paragraph where the fitted model is integrated with the expression analysis stage.

Following face detection, a plethora of image processing operations are carried out mainly for the purpose of normalising or otherwise aligning detected face regions or fitted models. The
operations include smoothing using various kernels, colour space conversion, as well as image rescaling, rotation and cropping. The parameters of these operations are controlled through a comparison of the detected face or model and a stored ideal version for expression detection. In [38], Chew et al. use the ‘constrained local model’ (CLM) developed by Saragih et al. in [39]. The CLM fits a face model to a primitive image region identified as a potential face. Based on local optimisation, the model’s vertices are picked to lie at certain pixel locations based on a probabilistic framework as shown in Figure 3. The vertices are then shifted individually as well as a group to best fit the stored model. This is an example of a computationally intensive pre-processing step, resulting in restrictions to offline applications. Nowadays, with an eye ultimately focused towards real-time applications, researchers prefer to do away with stochastic schemes throughout the classification pipeline to improve the overall performance. Preference is given to appearance-based algorithms which do require well-aligned faces. However, with speedy detection becoming a standard requirement, combining coarse face alignment with a large database (intrinsically provides averaging effects) and feature vector dimensionality reduction techniques a compromise can be achieved.

![Figure 3: ‘Constrained Local Model’ fitting scheme; an exhaustive local search for feature locations is followed by a stochastic optimisation strategy to maximise the responses (or likelihood) of one of the Point Distribution Models on the right [39]](image)

Once a face region/model has been located and appropriate pre-processing steps are carried out, the grand approach selected by the designer of each facial expression classifier comes into play. The grand approach or basis of the algorithm can usually be pigeon-holed into one of two categories: appearance-based or model-based. This dichotomy is common for many classification tasks in computer vision (e.g. rigid object recognition [40] and human pose tracking [41]).
The two expression classification approaches share many of the same steps; both require a training database, image processing, feature extraction, classifier evaluation, and a space to visualise or map extracted expressions (or their inferred emotions). The differences lie in the structures that represent each facial expression. In both approaches, the data obtained from the feature extraction stage is used in the classifier to produce a label for the input. In appearance-based approaches, the classifier’s result directly speaks to the existence or absence of each facial expression. On the other hand, model-based approaches are predicated upon a representative abstract model or framework that may subsequently infer facial expressions. More often than not, the model is constructed upon observed phenomena, for example, the detection of a combination of raised cheekbones and upward curving, parted lips represents an expression strongly associated with happiness. Thus, in model-based approaches, the classifier results infer the existence or absence of components of the abstract framework, while appearance-based approaches allow the raw data to directly infer the expressions in the input image without going through an artificial layer.

The most common abstract model used in today’s research is the Facial Action Coding System (FACS) first discussed by Ekman and Friesen in 1976 [42] and presented formally in 1978 [43]. A list of common facial movements known as Action Units (AUs), examples of which are shown in Table 1, are extracted from an image or sequence of images from the ‘CK+ Database’ [44].

Once the AUs present in an image are collected, their interpretation is up to the user. Ekman and Rosenberg [45] discuss the merits of some existing combinational understandings such as Ekman and Friesen’s early ‘Emotion Facial Action Coding System (EMFACS)’ [46] and its sequel, ‘Facial Action Coding System Affect Interpretation Dictionary (FACSAID)’ [47]. However, many current algorithms that utilize FACS do not attempt to make the interpretive connection from AUs to a labelled facial expression (or underlying affective state). Instead, papers such as [48], [49], and [50] leave the facial expression problem open for further investigation, satisfied with a detection scheme solely for action units.

A popular tool used in FACS based algorithms is the ‘Cohn-Kanade AU-Coded Database (CK+)’ [44] (discussed in Chapter V). The CK+ set was manually labelled, an extremely labour intensive process as mentioned by Chew et al. in [48]. In an interesting reversal, many of the algorithms that use the CK+ set are aimed at automating the tedious labelling process.
A common design choice made by researchers in this field is to pair model fitting during face detection with the model-based approach to classifying expressions. Tsalakanidou and Malassiotis, [52], utilised Lanitis et al.’s [53] ‘Active Shape Model (ASM)’-based algorithm for face localisation. A ‘Point Distribution Model (PDM)’ is fit to the detected faces as shown on the left in Figure 4. Tsalakanidou and Malassiotis then carry out geometric operations including point to point distance measurement, polygon area calculation as well tangential and normal vector derivation. These geometric values not only serve as the parameters of the ASM, but as the variables for detecting facial actions. A similar translation of these values to underlying emotions was developed by the authors, relying on logic and arithmetic rule-based mapping that also uncovers the equivalent AUs. Figure 4 also shows an example of a face with superimposed geometric measurements.

Table 1: Action Units in the Facial Action Coding System [43] (Table adapted from [51])

<table>
<thead>
<tr>
<th>AU</th>
<th>Description</th>
<th>Example Image</th>
<th>AU</th>
<th>Description</th>
<th>Example Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Inner Brow Raiser</td>
<td></td>
<td>22</td>
<td>Lip Funneler</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Outer Brow Raiser</td>
<td></td>
<td>18</td>
<td>Lip Puckerer</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Lip Corner Depressor</td>
<td></td>
<td>43</td>
<td>Eyes Closed</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Point Distribution Model (left), Superimposed Geometric Measurements for the Active Shape Model (right) [53]
Raouzaio et al. produced a custom model-based expression recognition framework in [54], similar to that of Tsalakanidou and Malassiotis [52]. However, Raouzaio’s method exposes one of the biggest weaknesses that many model-based approaches possess, not only within this research topic, but across the computer vision field: manual initialisation of a model. Initialisation is usually essential during the training stage; however, this does not take away from its chore-like, repetitive nature requiring human intervention. Thus, to obtain a large, well constructed training database, a high number of man hours are required not only for labelling the data, but also for verifying the result of this error prone task. Requiring manual initialisation in the testing (or classification) phase is usually a make or break point for an algorithm. The initialisation stage invariably acts as the processing bottle neck, especially in time sensitive applications.

Essa and Pentland’s [31] criticism of current model-based approaches like FACS leans heavily on the diversity of the human facial expression repertoire. While maintaining that some universality in expressions may exist, they claim that FACS fails to capture the subtle variations in expressions due to the ‘localised’ nature of AUs. Instead, Essa and Pentland claim that each facial expression is a collective set of actions, taking place across the entire visage in a cohesive manner, rendering them difficult to break down in a component fashion.

This criticism serves as an ideal segue to the discussion of appearance-based facial expression classification algorithms since parameterisation in the form of AUs or geometric measurements is replaced with the full facial image, satisfying Essa and Pentland’s locality qualms. Interestingly enough, appearance-based algorithms serve as a precursory step in many model-based approaches to expression classification, being used for extracting the parameters of the face model.

In practice, appearance-based approaches to extracting facial expressions can vary greatly, with differences occurring right off the bat in the image normalisation technique, through to the feature set extracted, and finally, with the machine learning classifier used to evaluate the features following the training stage. Regardless of the internal variations, these steps represent the standard itinerary for facial expression classification with this approach.

As mentioned earlier, normalisation methods in today’s algorithms need to be efficient, thus, stochastic optimisation for alignment is generally avoided. Arithmetic image processing operations such as rescaling based on eye position [55], greyscale conversion and histogram equalisation [50] do not require more than the four simple mathematical operators, and can make use of parallelised computer architecture such as Graphical Processing Units (GPUs).
The role of feature analysis in appearance-based approaches is to transform the underlying image data from the pixel shade space (colour or greyscale) to one that better partitions the various expressions. Feature analysis results in a feature vector containing a spatially consistent descriptor for each pixel or image region.

The features available for researchers in this field are all those used throughout the computer vision community. Local binary patterns [56], gradient based edges [57], Gabor filters [48], optical flow [31] and Haar wavelets [58] are some examples of image processing tools used to create feature vectors that are used to classify images. Multi-resolution analysis and sequential or frame-by-frame feature construction (includes all temporally-dependent features) are examples of extensions of these tools. These extensions may result in a mismatch in dimensions of the feature vector and those of the original image, thus, the transformation from the image space to the feature space may not be unique. The lack of uniqueness is a favourable characteristic when images containing similar facial expressions are mapped to the same point in the feature space, effectively creating expression clusters.

Following feature analysis, either training or evaluation of the classifier takes place. The classifiers available to researchers are as abundant in numbers as the features they use, with no shortage of examples for each type. The most basic form of any classifier places samples under one of two labels, where training is carried out with positive and negative samples in a binary scheme. Splitting the task of classifying a range of facial expressions, with varying intensities, into a set of binary problems is a viable option. However, a more complex classifier capable of multi-class labelling is more commonly employed.

Support vector machines (SVMs) have garnered the lion’s share of the attention in this field, as seen in [59], [60], and [61] due to the simplicity of their training and implementation and easily tuned parameters, namely, the kernel function and its contents. Artificial neural networks (ANNs) in all their varieties have been popular classifiers, as seen in [62], [63], [64], [65], and [66], partially due to their intrinsic multi-label classification capabilities. Decision trees [67], naïve Bayes classifiers [68], and hidden Markov Models [69] have also been implemented for facial expression recognition; however, the current state of the art results have come from SVMs.

A comprehensive summary of existing expression recognition methods up to 2009 is contained in Zeng et al.’s survey [70]. This survey does not limit the investigation to methods employing direct visual analysis of the subject, but rather extends to all media mode manipulation systems aiming at determining the subject’s affective state. Tian, Kanade and Cohn reviewed the state of facial expression research in their 2011 summary, focusing on state-of-the-art visual methods [71].
IV Existing Facial Expression and Emotion Classification Models

The topic of facial expression and emotion classification has included a number of well established and heavily researched methods. When considering classification techniques for said two characteristics, one cannot help but treat facial expressions as the outer shell of the underlying emotive states. Thus, it is not uncommon to amalgamate their classification medium, linking the visible facial expressions directly to a counterpart in the invisible emotion domain (e.g. smiling with happiness or frowning with anger).5

Representations of emotions and expressions in existing algorithms can be generally classified into dimensional approaches and categorical (or appraisal) approaches, however, arguments can be made with regards to the former being a manipulation of the latter.

Much of the work aimed towards developing a structured space for facial expression/emotive state comes from the realm of computer graphic reanimation and model deformation. Such details have become integral to the success of any computer generated form of entertainment, for example, in animated feature films, video games [72], [73], [74] and even more so for human computer interaction (HCI) [75], [76].

Dimensional approaches attempt to construct a space, orthogonal in some cases, within which expressions can be plotted in some continuous form with respect to an intensity value. The space is usually learnt in the training stage and subsequently used to classify test samples. In the relevant literature, facial expression usually infers underlying affective states (or emotions), thus some of the published dimensional spaces classify emotions rather than expressions although the same concept of intensity still applies.

In contrast, categorical approaches try to break down the multiple factors that construct and contribute to a human emotion. Cause and effect models are popularly implemented here, incorporating the environment within which the human (or virtual avatar) exists. This environment can include other people, natural phenomena and multiple sensual media such as sound and smell.

5 It is important to clarify that these frameworks are independent of the model-based vs. appearance-based approaches to algorithm construction explained in the literature review section
The most widely discussed and implemented examples for each approach are summarised below, including the motivation behind their existence as well their applicability to this project.

IV-1 Pleasure-Arousal-Dominance (PAD) [77]

Mehrabian and Russell developed the PAD model in an effort to bridge emotion research from the realm of social science to the more analytic, natural counterpart. Mehrabian's background as both an engineer and a psychologist drove the methodology behind building the PAD space (analytical experimentation), and the long term reasoning for its existence (better assessment of psychological disorders).

The declared aim of the PAD space is to provide 'a few basic dimensions suitable for the analysis' of emotion research problems. Given that dedicated non-clinical attention towards this topic has only picked up steam in the last 15 years, psychologists were the primary researchers in the field, predominantly examining the human angle of the subject.

Mehrabian’s PAD emotion model was developed over a series of studies that brought about three nearly independent scales, Pleasure-Displeasure (P), Arousal-Nonarousal (A), and Dominance-Submissiveness (D). Participants were required to grade images and situations (described in words) using antagonistic word pairs on a 16 level scale for P, and a 9 level scale for A and D. Subsequent statistical analysis of the polling results allowed for various emotional states to be plotted in a normalised space.

Significant research into the sufficiency of three scales (or dimensions) for the modelling of all emotions discernable by a human expert started as early as 1957 with Osgood, Suci and Tannenbaum’s inquisitive work in ‘The Measurement of Meaning’ [78]. It wasn’t until 1974 that Mehrabian and Russell began developing the early version of the PAD space [79], beginning with three disjoint scales similar to each of the three axes in the finalised framework. The need for a three dimensional space was corroborated by Shaver et al. in [80], where they suggested an alternative coordinate system with axes labelled as ‘Evaluation, Potency and Activity’.

Mehrabian pushed forth with the PAD scheme with attempts to classify higher order personality attributes and disorders. Up until Mehrabian's foray, personalities were classified through reliance on worded definitions, or categorical descriptions of various personality traits stemming from observational research. As Barrick and Mount summarised in their heavily cited report [81], the ‘big five’ personality model grew out of Raymond Cattell's extensive research in the 1940s that initially resulted in ‘16 primary factors and 8 second-order factors’ [82]. Cattell's factors underwent
repeated examination and dimensionality reduction, over the following 15 years, with Tupes and Christal’s 1961 report [83] surmising that a 5 factor model would be sufficiently comprehensive. The factors decided upon in [83] differed from the ones commonly seen in literature from the 1970s onwards, which were described by Warren Norman in his 1963 work [84]. Norman’s breakdown of personality factors included ratings of ‘Extroversion, Emotional Stability, Agreeableness, Conscientiousness, and Culture’. These labelled factors became known as the ‘Norman’s Big Five’.

Following extensive surveying and experimental studies, Mehrabian used regression analysis to relate the PAD space to existing personality description frameworks, such as Goldberg’s ‘Markers for the Big-Five Factor Structure’ [85] and Jackson’s ‘Personality Research Form Manual’ [86]. Mehrabian succeeded in finding relatively reliable coefficients for each of the three dimensions, yielding equations that describe various personality types, thus providing a previously absent precedent to analytical research in this field.

IV-2 Ortony, Clore and Collins (OCC) Model of Emotion [87]

The OCC model is a good example of a categorical classification scheme. Although the model incorporates both qualitative and quantitative aspects, the former is commonly implemented for its hierarchical structure, wide range of labelled emotions and logical approach to determining underlying emotive states.

Ortony, Clore and Collins published their framework in the 1988 book ‘The Cognitive Structure of Emotions’ [87] where they began with the goal of ‘specifying the global structure interrelating different emotions as well as the characteristics of individual emotions’. For an abstract pursuit such as this, the authors aimed to construct a logical based rule book while maintaining a theoretical psychological undertone.

The OCC model is governed by ‘valenced reactions’, with the aforementioned cause and effect relation sitting at the forefront of the emotion suggestion scheme. The reactions controlling the resultant perceived emotion involve cognitive evaluation of three criteria within each scenario:

1. Consequences of Events: Being pleased or displeased with the outcome or presence of an event with respect to self fortunes or those of others.

2. Actions of Agents: Approving or disapproving of self as well as others’ actions, how they reflect upon their portrayer, and how they link up with concurrent events.
3. Aspects of Objects: The affection for objects within the evaluation domain, how much they ‘appeal’ to the subject.

The authors’ motivation for this structure is that emotion formation is a collaborative effort between events, agents, and objects in association with the subject’s internal goals, standards, and attitudes, respectively. Thus, the same scenario could stir different emotions within different individuals, in line with common intuition.

Within the author’s reasoning for the existence of emotion theory, they mention the ambiguity of the ‘language of emotions’. Their framework lists a total of 22 emotions that could be experienced; however, they acknowledge that this listing of words describing distinct emotions may not apply to non-English speakers. This conflict is resolved through the authors’ sub-goal of describing emotions ‘in as language-neutral a manner as possible’. Thus, in the purest form, emotions would simply be described in their relation to the criteria evaluation listed above, with no worded label attached.

The OCC model has been implemented in a variety of applications; however, it has mostly appeared in HCI settings within the realm of computer science [88], [74]. The lack of a strict expression intensity calculation methodology renders the model quite dependant on supporting frameworks, such as that used in Gebhard’s Layered Model of Affect (ALMA) [89]. Here the author combines the verbal descriptors of the OCC model with the dimensionality of Mehrabian’s PAD space for the purpose of creating a personality-mood-emotion, temporally dependant application.

IV-3 Applicability to this Project

The fact that this project is concerned with an analytically minded classification of facial expressions in still images and video should not entirely eliminate either approach from consideration. However, it is difficult to ignore the merits of implementing a dimensional space for mapping emotions. Several integral stages or components of our proposed algorithm lend themselves to such a space. Starting with the selected training database (BU-3DFE [90]) and the underlying emotion model (Ekman’s ‘big six’ [2]), we are presented with a set of dimensions (or axes), one for each emotion. The classifier picked for this project, a random forest (RF) evaluator followed by an SVM or k-NN (k-Nearest Neighbours) labeller, allows for statistical collection of evaluated results as opposed to a singular label for each test case. Thus, near-continuous values
or scores for each emotional dimension are achievable, with the granularity of the scale being controlled by the number of trees within each forest.

Due to the readily available labels in the used databases, a transformation into a different space, such as PAD [77] or the more dated Frijda dimensions [91] could eventually be achieved. Furthermore, the six labels (seven including neutral) that Ekman proposed are quite distinct in their nature, resulting in seemingly independent dimensions. Investigating the correlation, or lack thereof, between these dimensions is one goal of this project as discussed in Chapter IX.
V Databases

Data required to train and test an unconstrained facial expression recognition system must address all aspects of the task at hand. Given the proprietary nature and strong reliability of the face detector employed, the focus of the database evaluation stage is the applicability to the expression recognition portion of the project. The ideal database would combine variations in head pose and subject demographics with labelled facial expression intensity pertaining to a set of predefined emotions.

A number of databases were examined for their relevance and each was summarised with respect to their contents; including participants, labelling scheme, and comprehensiveness. The databases included in this list were either used in the project or reviewed for their significance in the field. A number of other facial expression research databases not used or reviewed here are summarised in [71], [92], and [93].

As mentioned in Chapter II, the expressions displayed by humans are true and real only when they reflect the internal affective state of the person. This point is of serious consequence when constructing any recognition system that relies on a database of staged scenarios to mimic real life facial expressions. A database consisting purely of true emotions while providing comprehensive coverage of all other variables involved in an unconstrained facial expression recognition system is not presently available.

An assumption that true or natural expressions are all similar is necessarily made for this project to be theoretically sound and in line with Ekman’s universality of expressions doctrine. The implications of using a staged database become evident in the testing stage where the false expression in the training set may differ from another ‘false’ expression in a staged test set and even ‘true’ expressions in a ‘natural’ test set.

Unfortunately, due to the aforementioned unavailability of a comprehensive natural database, the BU-3DFE database was seen as the next best alternative. The graded intensity level and large number of participants is presumed to counteract some of the disadvantages of using a staged database for training the expression classifier.
The BU-3DFE database contains static three dimensional Virtual Reality Modeling Language (VRML) face models along with their respective texture files. Each individual was instructed to display four intensity levels pertaining to each of the six Ekman expressions (anger, disgust, fear, happiness, sadness and surprise), as well as a single expression representing a neutral affective state. A high-density 3D point cloud was captured for each model, along with a pair of images taken at 45° deviations either side of the facial midline. The images were used to create a texture map for the point cloud. Using projection of the 3D point cloud and the corresponding facial texture, multiple head poses can be generated for all seven expression sets.

An impressive number of participants was reported (56 females, 44 males), with a wide age range (18 to 70) and a large collection of racial backgrounds (‘White, Black, East-Asian, Middle-east Asian, Indian, and Hispanic Latino’). The heavily constrained lighting used to capture the point clouds and texture maps allows for the artificial generation of multiple illumination models, which can be applied to the original captured data. Thus, as a training set, this database serves the requirements of an expression recognition system, covering several integral variables: expression intensity, head pose, illumination and subject demographics.

An extended breakdown of training dataset generation extracted from the BU-3DFE database is provided in Section VII-1.

The KDEF database was initially developed for neuroscience research; however, it has since been used in the computer vision field due to its applicability. A total of 70 participants displayed the 6 Ekman expressions in addition to neutrality, at 5 different head yaw variations to produce a total of 2,450 images. The experiment was then repeated, bringing the total number of images to 4,900.

The authors kept an eye out for several important issues as the database was developed. The environment (lighting, background, camera distance) was kept constant throughout the capture process and the subjects were asked to remove all accessories (hats, glasses, facial hair and make-up).
The RaFD provides multi-pose, multi-expression, multi-gaze images displayed by a total of 67 models (25 females, 42 males). This dataset is quite unique in its inclusion of children as subjects, in addition to adults of various ages, thus, providing a good platform for evaluating the performance of a system trained solely on adult faces\(^6\).

The variation in pose consists of the subject being captured simultaneously by 5 different cameras at yaw angles -90°, -45°, 0°, -45°, 90°. Ekman’s six expressions, in addition to neutral and contemptuous expressions were displayed by the subjects after receiving specific instructions from a certified FACS expert. Finally, each expression pose combination was captured with the subject gazing straight ahead, to their right and then to their left. The database totalled to 8040 images.

In addition, the authors provided validation data obtained from a surveying experiment taken by 276 university students. A five point scale was used to quantify opinions of the frontal image with respect to, amongst others, the intensity, clarity and genuineness of the depicted expression. This experiment provided good insight into the dataset’s characteristics as summarised in [95].

The JAFFE database contains 216 greyscale images of 10 Japanese female models, each exhibiting seven facial expressions representative of six emotions (anger, disgust, fear, happiness, sadness and surprise) and a neutral affective state. The models display between 3 and 5 intensities for each emotion, and the resultant images were subsequently graded on an ascending scale of intensity ranging from 1 to 5 for each emotion.

An average intensity grade for each of the six emotions was supplied for every image through psychological evaluation carried out by a group of 60 Japanese female students on the images. This semantic grading scheme was useful as it provided a preliminary six dimensional coordinate for each image. An interesting addendum was the removal of the fear emotion grade, as well as the images displaying fear, for two reasons. Firstly, the authors believed that the models had trouble simulating fear convincingly. Secondly, they mention scientific evidence claiming that fear

\(^{6}\) The youngest subject in the BU-3DFE was 18 years old
is in fact processed differently from the remaining emotions; however, no reference is given to support this claim.

Overall, this database could be useful for preliminary training or testing, particularly due to the semantic ratings provided through the author’s surveying experiments. However, the narrowness of model demographics, invariable illumination, and singular (frontal) pose in all constituent images limits the utility of the database. Thus, the JAFFE database was simply tested using the facial expression framework constructed for the project.

V-5 Psychological Image Collection at Stirling – Pain (PICS - pain) [97]
Testing – Still Images

This image set is geared towards covering a wide range of facial expressions displayed by humans while experiencing pain. The 599 images contained in this set come from 13 females and 10 males, where, on average, each subject displays two expressions representing each of Ekman’s six basic emotions, two representing neutrality, and ten representing painful affective states. Pose variations are also contained in this image set. However, the expression repertoire is made available at a frontal pose with minor variations in the yaw, roll and pitch angles.

V-6 Static Facial Expressions In the Wild (SFEW) Database [98]
Testing – Still Images

The SFEW database contains screen captures extracted from feature film video clips compiled for the AFEW database (see Section V-9). A total of 700 still images containing individual actors in a variety of poses make up the database. The actor’s portray one of the six Ekman emotions or a neutral facial expression.

The authors agree with the database trueness concerns expressed at the beginning of this chapter. They also added that movies provide ‘close to real world environments’ that improve on the somewhat constrained conditions under which most databases mentioned in this chapter were captured.

V-7 Binghamton University 3D Dynamic Facial Expression Database [90]
Testing – Videos

The BU-3DFE authors extended the data capture framework by adding a temporal progression of the captured VRML models. The same projection technique used to produce the training set images (see Section VII-1) was used to produce consecutive frames, later compiled into short video clips. The authors used an affective arc akin to the MMI database sequences, beginning
and ending the 3D progression with a neutral expression. Unlike the BU-3DFE database, the dynamic version contains a singular expression label relating to the apical frame of the progression without an additional intensity label.

Demographic variations amongst the 101 subjects are impressive, with each displaying a sequence for all six Ekman expressions, resulting in 606 3D model sequences.

V-8 Facial Expressions and Emotion Database (FEED) [99]
Testing – Videos

This database provides an ideal test platform for the implemented project. The author stresses the importance of avoiding ‘played’ emotions for the same reasons mentioned in the beginning of this chapter. Thus, a test bench aimed at stirring different emotions in the participants by showing them video clips was used on 18 individuals. The participants were recorded displaying 3 versions of each state (anger, disgust, fear, happiness, neutrality, sadness and surprise).

This method of obtaining ‘natural’ data makes the assumption that the video clips will stir the same emotion within each individual. However, the purity of the emotions is sullied by the subjects’ knowledge that they were being recorded. Nevertheless, the FEED database provides a rare structured dataset for evaluating unconstrained pseudo-spontaneous expressions.

V-9 Acted Facial Expressions In The Wild (AFEW) Database [100]
Testing – Videos

The AFEW contains short clips extracted from feature films where a single actor portrays a progression of facial expressions labelled as one of Ekman’s six (or neutral). As mentioned in the SFEW review in Section V-6, the authors’ motivation for constructing the AFEW set is that movies provide a more realistic environment than databases captured in laboratories.

The AFEW consists of 957 video clips, each of 300 to 5400 milliseconds in duration, covering said expressions with no clear breakdown of subject demographics. The heavily unconstrained conditions contained in the database make for a strong test platform.
CAS Pose, Expression, Accessory & Lighting (PEAL) Database [101]
Reviewed for Database’s Comprehensiveness

The PEAL database attempts to capture variations in pose (27 per subject), expression (6 per subject: neutral, smiling, frowning, surprised, eyes closed, mouth open), lighting (at least 9 per subject), accessories (6 per subject), and background (up to 4 per subject). Combinations of the different variables were not exhaustive; and pose and expression changes were varied together.

The full PEAL database contains 99,584 images of 1040 subjects (57.2% male). However, only a partial database (30,871 images) is available for research purposes. Pose (present in 21,840 images) and expression variations (present in 1,884 images) are well represented in the partial database.

The relevance of the PEAL database is on par with that of its counterparts, lack of expression intensity grading limits its applicability to emotion research, however, the full database promises a vast array of combinational arrangements, simulating unconstrained scenarios that are the core of the testing stage of this project.

CMU Multi-Pose, Illumination & Expression (Multi-PIE) Database [102]
Reviewed for Database’s Comprehensiveness

The CMU Multi-PIE database aims to cover the majority of pose and illumination variations while incorporating six basic facial expressions (neutral, smiling, squinting, surprised, disgusted and screaming). The total number of unique participants over four image capturing sessions was 337, with 264 of those recorded more than once. Demographics for this database included 69.7% males, a racial distribution consisting of 60% Euro-American, 35% Asian, and 5% other. The purchased database contains extended demographic information.

The physical imaging setup included 13 cameras arranged in a semi-circular arc facing the subject, along with 2 elevated cameras at the 45° deviation point (simulating surveillance applications) either side of the head. Controlled illumination was achieved through the use of 18 flashes, distributed in a similar fashion to the cameras and providing 19 different lighting conditions. During each recording session, two or three facial expressions (always including neutral) were modelled by the participants. Permuting through the number of sessions, participants, cameras, illumination and facial expressions, resulted in a total of 755,370 images.

The multidimensionality and expanse of the CMU Multi-PIE database is quite useful for illumination and pose investigation, two areas integral to any emotion recognition system;
however, as an expression-centric database, CMU Multi-PIE falls short due to the single expression intensity displayed.

V-12  MMI Facial Expression Database [103], [104]
Reviewed for Database’s Comprehensiveness

In 2002, Maja Pantic, Michel Valstar and Ioannis Patras started the MMI database, stressing the importance of both static image and motion sequences in emotion research. According to the authors’ analysis of existing face and facial expression databases, a ‘glaring’ lack of easily searchable, well documented options was found.

The MMI database is continuously growing, allowing for new additions through submission to the system administrators. The initial database (or Part I) consists of 740 static images, and 780 motion sequences displaying instructed facial expressions representing individual or combinational action units as well as video progressions from neutral to expressive and back to neutral. The latter media was inspired by the CK+ database [44]. However, a finer temporal segmentation with respect to the labelling scheme was used, as well as the extension of the expression display sequence past the apex, back to a neutral affective state. Parts II and III consist of 238 higher resolution facial video sequences and still images respectively. In addition, six emotions (anger, disgust, fear, happiness, sadness, and surprise) plus neutral were artificially simulated by the participants in the first three parts, with variations in background (cluttered vs. solid colour), illumination (natural vs. controlled), accessories (glasses vs. no glasses) and pose (frontal vs. profile).

Certified FACS coders instructed the participants throughout the first three parts. However, for parts IV and V, stimuli expected to invoke certain emotions were used to produce facial expressions representative of happiness, surprise and disgust by the participants. Invoking the remaining emotions was seen to be unethical by the authors. As an analogy to the Heisenberg uncertainty principle, ‘spontaneous’ sequences are some of the most difficult to capture, always possessing a level of artificiality unless the subjects are unaware of the camera’s presence.

A concise breakdown of the full database’s participant demographics was not clearly discussed, although for Parts IV and V (arguably the author’s largest contribution to the facial expression research community), a summary is provided. Participants for the spontaneous sequences included variation in gender (12 female, 13 male), narrows age difference (20 to 32) and race (10 European, 3 South American, and 12 Asian).
Spontaneous sequences could be considered in either the training or test stage and in an ideal setting they would be used for both. However, the present day relative rarity of these sequences results in their use as testing material reflecting the eventual real world scenarios that expression recognition systems aim to classify.

**V-13 Cohn-Kanade AU-Coded Facial Expression Database 2 (CK+) [44]**
Reviewed for FACS relevance

In 2000, Cohn, Kanade and Tian [105] assembled a database containing FACS coded expression sequences that was used extensively in emotion related research. In 2010, the database was extended from 486 to 593 sequences performed by 123 subjects. The participants were distributed amongst gender (69% female), race (81% Euro-American, 13% Afro-American and 6% others), and age (18 to 50 years old). Furthermore, two views are provided in the database, frontal and 30° deviation from the facial midline.

The sequences progressed from a neutral expression to an apical expression representative of the requested emotion. Action unit (AU) labels were provided for the final frame of each sequence. Both the AU labels as well as the resultant modelled expression were verified against the original requested emotion.

In addition to the ‘posed’ expressions, the authors added 122 spontaneous smile sequences obtained between scheduled emotion modelling requests. These ‘non-posed’ sequences are quite rarely found in databases due to privacy concerns related to filming unsuspecting individuals, as well as uncertainty with respect to the actual expressed emotion.

Overall, the CK+ database provides a well constructed training and testing framework for a FACS based system. Benchmarks for AU detection are provided by the authors using SVMs, in addition to values reported by third parties.
Table 2 summarises the databases used for training and testing in this project.

<table>
<thead>
<tr>
<th>Name</th>
<th>Media Type</th>
<th>No. Unique Subjects</th>
<th>Relevant Expressions</th>
<th>Number of Data Elements Used</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>BU-3DFE</td>
<td>3D VRML Models</td>
<td>56 F + 44 M</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>100 Models (See Section VII-1)</td>
<td>Training</td>
</tr>
<tr>
<td>KDEF</td>
<td>Still Images</td>
<td>35 F + 35 M</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>4900 Images</td>
<td>Testing</td>
</tr>
<tr>
<td>RaFD</td>
<td>Still Images</td>
<td>25 F + 42 M</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>8040 Images</td>
<td>Testing</td>
</tr>
<tr>
<td>JAFFE</td>
<td>Still Images</td>
<td>10 F</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>213 Images</td>
<td>Testing</td>
</tr>
<tr>
<td>PICS – pain</td>
<td>Still Images</td>
<td>13 F + 10 M</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>316 Images</td>
<td>Testing</td>
</tr>
<tr>
<td>SFEW</td>
<td>Still Images</td>
<td>Not defined</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>700 Images</td>
<td>Testing</td>
</tr>
<tr>
<td>BU-4DFE</td>
<td>Sequential 3D VRML Models</td>
<td>58 F + 43 M</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>707 Videos</td>
<td>Testing</td>
</tr>
<tr>
<td>FEED</td>
<td>Videos</td>
<td>9 F + 9 M</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>378 Videos</td>
<td>Testing</td>
</tr>
<tr>
<td>AFEW</td>
<td>Videos</td>
<td>Not defined</td>
<td>AN, DI, FE, HA, NE, SA, SU</td>
<td>900 Videos</td>
<td>Testing</td>
</tr>
</tbody>
</table>

Table 2: Summary of databases as used in this project

7 F = Female, M = Male, AN = Anger, DI = Disgust, FE = Fear, HA = Happiness, NE = Neutrality, SA = Sadness, SU = Surprise
VI Face Detection

Face detection as a computer vision problem has historically been approached with an interest in one facial characteristic at a time, while ignoring the remaining facets. Examples of this observation include Sung and Poggio’s ‘vertical frontal’ restriction in [106], as well as Manjunath’s et al. in [107] where low noise and no background clutter are required. The artificiality of these partial problems highlights the complexity of the fully unconstrained problem. Listed below are commonly encountered complications and their respective solutions sampled from the vast ocean of the existing face detection literature.

VI-1 Topics of Concern in Face Detection

Lighting

Varying illumination effects are some of the most difficult to resolve given that little to no a priori information is known about the direction, intensity, type or colour of the light in an unconstrained environment. Another issue common to scenes containing discontinuous illumination is the limitation of the physical sensor used to capture the scene.

A commonly employed technique used to circumvent varying illumination across an individual image or frame is histogram equalisation. Histogram equalisation attempts to better distribute the colour space to accommodate varying lighting exposure.

Unfortunately, contrast improvement through such histogram flattening generally decreases the overall brightness of the image. Ibrahim & Kong presented an algorithm in [108] that maintains brightness while improving the contrast through adaptive equalization. The algorithm operates entirely in the histogram space, partitioning the range of values following one dimensional Gaussian smoothing. The partitions are determined based on the maxima of the smoothed histogram, whereby dense regions in the histogram are stretched across relatively larger ranges and pixels are assigned with their new values. Finally, brightness is maintained by changing the mean of the manipulated histogram to resemble that of the original image.

Pose

Head pose in the real world can be defined as a three dimensional state consisting of rotations about three independent axes centered at the middle of the head. Once projected onto a two dimensional image plane, the inverse problem of extracting actual head pose from an image
could be a well-defined problem in the strictest sense of the term. However, it can also be seen to be undetermined.

Close-by poses, such as \((\text{roll}, \text{pitch}, \text{yaw}) = (0^\circ, 0^\circ, 0^\circ)\) and \((0^\circ, 0^\circ, 10^\circ)\) should map to adjacent points in the image space, furthermore, no pose should be inferred by more than one image appearance. The final requirement, although trivial, is that every detected face in the image should be paired with a pose.

Accessories

A good measure of any algorithm’s robustness is the ability to accommodate previously unencountered additions to the data it was trained on. Facial accessories can vary greatly in shape, size as well as type. Depending on the training and test sets employed, accessories could include facial hair, jewellery, eyeglasses, or hats. Whether these accessories are considered part of the face or as external objects [109] is dependent on the algorithm used.

Model based algorithms that explicitly identify accessories could benefit from the information gained when their presence is detected such as [110] and [111] where glasses and beards, respectively, are removed to show an underlying face. However, as with most model based techniques, an extensive database list of potential cases is possible, so that these types of algorithms are usually well suited only to heavily constrained scenarios.

A good example of an appearance based algorithm accommodating various accessories while maintaining high detection rates is contained in the seminal work of Turk and Pentland [112]. Principal component analysis carried out on the training set’s feature vector matrix was used to reduce the dimensionality of the original feature space while maximising the variance within the new space. Using a training set predominantly consisting of accessory free faces interspersed with some sporting facial hair, a successful facial detection base function was developed. The algorithm was able to overcome the presence of foreign accessories such as eyeglasses through the metric used to distinguish faces from non-faces. The distance measure used is the principal component weights extracted from each test image, which by definition will be concerned with the similarity to the most popular and consistently present aspects of the training set. This measure intrinsically suppresses the effect of minor changes in facial appearance imposed by accessories.
Demographics

A number of other sources of difficulty include skin colour [113], race [114], age [115] and gender [116]. Humans possess a documented 'cross-race recognition deficit' [117], while pinpointing age and gender pose challenges of their own. Thus it would be impractical to expect a face detection algorithm to differentiate these characteristics with perfection, although, in general their variation should not impede successful detection.

Automation and Real-time Performance

Extending effective face detection methods to real-time performance has become a staple requirement for practical applications in recent years. Furthermore, some face detectors require manual initialisation for face tracking, an undesirable step in most practical applications. Thus, for a face detector to be truly prepared for the scenarios listed in the sections above, while maintaining readiness for practical tasks, both speed and independence are needed. The Viola and Jones [34] face detection framework is a prime example of a fast, self-contained system that maintains its promised rates even on low performance machines.

VI-2 Requirements of Unconstrained Face Detection

Unconstrained face detection aims to correctly localise faces in a given image or sequence of images, under a wide range of non-extreme viewing conditions. From the summary above, it is clear that an all-encompassing detection system, mimicking human abilities would not only be excessive at this time, but also rather unattainable given the lack of cognitive abilities in computer applications. Thus, it is important to properly define these non-extreme viewing conditions. In addition to accommodating demographic differences, the proposed algorithm should be robust to variable lighting, head pose and accessories.

We assume that the ambient illumination levels in the image or sequence being considered are allowed to fluctuate from frame to frame, while directional lighting does not dramatically alter the appearance of the faces in view. Furthermore, the colour of the light in the scene should not pose a problem. One proposed technique used to avoid common visual spectrum (VS) lighting problems is to utilise infra red (IR) properties [118]. Although the issues discussed earlier regarding VS illumination apply to IR, heat emission from mammalian bodies is quite consistent under a wide range VS levels [119] and can be of great use in unconstrained environments. However, the relative rareness of IR cameras has restricted their use to surveillance geared applications.
Head pose is a more measureable quantity than directional illumination, and subsequently shadows, in the image domain given that most face detection algorithms are trained on a discrete number of labelled angular variations in head position with respect to the frame at hand. Intermediary head poses can be calculated using regressive calculations given the confidence levels or probability density function of the discretised range [120].

Accessories, including facial hair, can be considered non-constant features of a face [121] (mirroring the insight of Turk and Pentland [112]). The treatment of accessories as partial occlusions to faces in the scene is therefore intuitively acceptable [122] [123]. Both appearance and feature based face detection methods require a streamlined occlusion absorption technique. Heiseley et al. [124] claim that a component-wise breakdown of the image, such as gridding, is essential to proper classification in the presence of partial occlusions whereby the analysis of a concatenation of individual fragments produces a more meaningful confidence score.

VI-3 Approaches to Unconstrained Face Detection

As Zhang and Zhang [125] mentioned in their 2010 report, processing speed and memory concerns have almost disappeared, leading to a drift towards appearance based approaches to computer vision problems. Appearance based algorithms are those where labelling is carried out based on the relation between features extracted from the candidate image, and those obtained from training images used to construct a machine learning classifier. Examples of classifiers used in appearance based algorithms include neural networks [126] [127], support vector machines with neural networks [127], and Fisher discriminant analysis [128]. Features in appearance based methods are usually pixel neighbourhood characteristics such as intensity gradients (edges) or logical relationships (LBP). A more detailed description of features and learning tools used in computer vision, and more specifically for expression recognition is provided in Chapter III.

In 2002, Yang et al. [129] identified three other doctrines for the automated analysis of human faces. Firstly, ‘knowledge based’ methods try to emulate human understanding in its cognitive form. Exhaustive logic rulebook construction is required here; however, it is an arduous task, particularly in a non-deterministic, exception filled case such as this one.

Secondly, ‘invariant feature’ approaches usually carry out similar analysis steps as appearance based methods, however, the former focus on locating supposedly consistent aspects of human faces as opposed to the more holistic analysis carried out in the latter methods. Examples of ‘invariant features’ include skin (well surveyed in [130]) and eyes (well surveyed in [131]) in
addition to hairline, cheekbone, mouth and nose localisation. Higher level statistical models are constructed for individual, or groups of, detected facial features and the confidence scores as well as some inter-model relations become the classifying tool. The goal of part by part identification is to reduce the computational cost of locating an entire face. Variation within each aspect of the face is meant to be captured by its detector, effectively dividing up the dimension of the full facial feature space. Furthermore, by ignoring external factors for each detector, each divvied up space is effectively reduced.

Thirdly, ‘template matching’ methods use a facial model constructed by the user where geometric relations between features are used. These relations include relative size, distance, and presence of features, with the similarity (or deviation) of an image from said relations determining whether or not it is classified as a face. Template methods require subsequent optimisation steps to obtain an effective, generalised application.

Overall, ‘invariant feature’ techniques have achieved favourable results particularly at the turn of the 21st century. However, as mentioned earlier, the improvement of computational power since then has allowed for more comprehensive classification developments using ‘appearance based’ methods. Training of ‘appearance based’ methods takes place on large sets of labelled images where the subjects in the pictures are instructed to strike particular poses and external factors such as illumination are varied in a controlled manner. Meanwhile, training of ‘invariant feature’ methods requires manual segmentation of the same images, a tedious task that adds a chance of human error. More information is provided in the Chapter V.

The detector used in this project is the Pittsburgh Pattern Recognition (PittPatt) package, a proprietary SDK providing a number of face detection and tracking tools. PittPatt adheres to the requirements set out above, with an impressive set of metadata provided for each detected face. A more detailed breakdown of the PittPatt system is given in the implementation portion of this thesis.

VI-4 Pittsburgh Pattern Recognition (PittPatt) [132]

The PittPatt face detection framework is a proprietary SDK containing a set of easily integrated tools providing data pertaining to faces in a single or sequence of images. The academic licence provided to the Visual Surveillance Group at McGill University permitted complete access to the PittPatt toolkit allowing for a highly customised detection stage. The unconstrained face detection summary was used as a criterion list when settling on the choice of PittPatt. With a few noted exceptions, PittPatt was found to be comprehensive in its detection and tracking abilities.
Burdensome problems such as demographic variations, partial occlusion, and the presence of accessories were no longer obstacles to successful detection. Furthermore, pose changes and multi-scale search were not only dealt with in a streamlined manner, metadata detailing pose (albeit on just two rotational axes), facial dimensions and particular landmark locations are explicitly returned.

Due to the proprietary nature of the PittPatt package, specifics regarding the analytical features used to locate faces and calculate the aforementioned peripheral information are not openly available. However, a summary of the package’s contents is compiled below, with some insight into the pertinence of certain results to the goal of expression recognition.

Finding faces in an image can take on a cascaded scale search format that analyses different regions and sizes of the image. This exhaustive search method is a departure from visual selection in humans where certain cues in the image focus the attention to particular image regions. The cues in human vision are obtained through parallelised multi-scale feature evaluators. Rather than employ a number of scale specific feature calculators as present in humans, many face detectors (including the insightful work by Fleuret and Geman [133]) maintain a fixed feature extractor size and analyse a sequence of resized versions of the original image in an effort to locate faces of differing scales. This technique permits a sort of foreground/background segmentation through the order in which faces are detected. PittPatt’s search pruning options emulate this segmentation method by restricting the search scales and aborting the search once a face of a certain size is found.

Furthermore, facial dimensions and centre location can be restricted in both absolute and relative fashions. Adaptive change in these parameters serves a larger purpose in tracking applications where estimates produced by filters can improve performance by ignoring unlikely scales and regions. These filters are defined by the percentage of the image that a potential face could cover, as well as restricting the location of the potential face’s center. A complete facial tracker and recognition toolkit is provided as part of PittPatt; however, the tracking algorithm used by the authors is not explained. Our assumption is that search space is varied from frame to frame according to the characteristics of the last detected face.

In addition to size and location, the metadata for each detected face includes a head pose estimate in the form of angular deviations around the Cartesian axes.

At the time of writing of this report, only two axial rotations (yaw and roll) were calculated, however, plans to include the third rotation (pitch) were mentioned in the documentation. The
centroid location, height, width and pose will determine the transformations necessary to normalise the detected face, whether in the training or the testing phase.

Physical feature locations can be optionally extracted from the calculated face, giving PittPatt great applicability to model-based algorithms. This extraction process requires additional computation, however, once found, the features contribute to a more accurate head pose estimate.

As mentioned in the methodology summary, a set of training images will be constructed using the BU-3DFE databases. This proposed arrangement involves a verification stage where the PittPatt framework will be used to confirm the dimensions and location of each face in the raw image set. Ultimately, the output obtained from PittPatt will contribute greatly to determining image pre-processing parameters and Random Forest classifier sets used.
VII Methodology and Implementation

This section provides a summary of the methodology used to detect and classify facial expressions in images and video sequences. A detailed examination of each subsection of the algorithm is provided further into this chapter and the overall processing flow is visualised in Flowchart 4 in Appendix II. The project consisted of five major stages:

1. An extensive image database was prepared for training the classifiers.
2. Feature analysis was carried out on the database contents.
3. Classifiers were trained on feature vectors extracted from the images in the database.
4. A temporal integration mechanism for accumulating and averaging classification results for individuals in a sequence of images or video stream.
5. A testing strategy was implemented to evaluate the performance of the system.

The BU-3DFE Database [11] was the singular source of facial expression training data. The three dimensional models contained in the database, along with their associated textures provided an artificial platform for producing a customised image set. The human subjects in the BU-3DFE database provided intrinsic variations in: race, age, and gender while the images generated from the database varied uniformly in:

a. Head Pose
b. Expression – Anger, Disgust, Fear, Happiness, Neutrality, Sadness and Surprise
c. Expression Intensity\(^8\) - 1 (Low), 2, 3, 4 (High)

Following raw image database generation, a face detector was used to localise the face in each training image. They were cropped, resized, pre-processed and analysed for various features using standard image processing techniques. Pre-processing of the training images consisted of colour space conversion, smoothing using a Gaussian filter, and illumination normalisation.

The results of the feature extraction step were compiled and the Principal Component Analysis (PCA) [134] dimensionality reduction technique was invoked to produce a more manageable descriptor or vector for each training image in the database. The reduced vectors were then used to train three collections of RF classifiers for each head pose, under the three different schemes. A labelling classifier (labeller) consisting of an SVM or k-NN to be paired with each RF was trained on the probabilistic output of the binary RF set.

\(^8\) Hereafter, ‘Expression Intensity’ will be referred to simply as ‘Intensity’
The temporal integration strategy was constructed to accumulate the output of the labeller when processing a video stream. Frames were processed individually while maintaining a running average for each individual face. The label attached to each face as the video progressed was based on the running average computed within a window surrounding each frame.

To evaluate the performance of the system the tools developed in the four previous stages were combined into a streamlined pipeline. The testing process began with the face detector providing location, size, landmarks and pose of faces in the test media (images or videos). The facial metadata controlled the cropping parameters and rescaling factors. Following face image pre-processing, steps mimicking those carried out in the training stage, that is, feature analysis and eigenspace projection, produced a vector ready for classification. Subsequently, the appropriate classifier set used for determining the expression of each face was selected according to the head pose detected. The probabilistic output of the RF collection was processed by the labeller to determine the label assigned to the current face (and combined with a running average in the case of a video stream).

VII-1 Training Database Preparation

Part 1: Raw Image Set (RIS)
Multi-Pose, Multi-Expression, Multi-Expression Intensity Image Generation from the BU-3DFE Database

The high resolution three dimensional (3D) VRML models contained in the BU-3DFE database [90] were used to generate sets of images for the eventual formation of a training set for the facial expression classifier. Mathworks’ Simulink 3D Animation Toolbox was used to manipulate the 3D VRML face models to achieve consistent variations over three different quantities:

1. Facial Pose
2. Facial Expression
3. Facial Expression Intensity

Examples of variations in each quantity are provided below. This unverified⁹ collection is known as the Raw Image Set (RIS).

---

⁹ Verification of the RIS is explained further on.
Pose Variation:

Variations in the pose were generated by incrementing the yaw angle of the 3D model from -90° to 90°, 15° at a time. Since PittPatt provides a value for the roll of detected faces, variations in this quantity were counteracted through simple in-plane image rotation required to return the face to 0° roll. Finally, the pitch was not varied as PittPatt does not provide data with respect to this quantity. Figure 6 shows the yaw variation for subject ‘F0003’.

Expression Variation:

Figure 7 shows subject ‘F0001’ displaying, in sequence, the six expressions representing anger, disgust, fear, happiness, sadness and surprise, all at maximum intensity, along with the neutral expression. Figure 8 shows subject ‘M0038’ displaying these same expressions.
Figure 8: Subject ‘M0038’ displaying expressions representing ‘anger, disgust, fear, happiness, sadness and surprise’ with maximal intensity, as well as the neutral expression

Expression Intensity Variation:

Figure 9 shows subject ‘F0001’ displaying all 4 intensities of anger, disgust, fear, happiness, sadness and surprise. Note that the neutral expression was only ever displayed at a single intensity, as shown in Figure 7 and Figure 8.
Figure 9: ‘F0001’ displaying expressions representing ‘anger, disgust, fear, happiness, sadness and surprise’ at increasing intensities from left to right
The RIS contains 32,500 images as per the breakdown shown in Table 3.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>100 [56 Female, 44 Male]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-neutral Expressions</td>
<td>6 [Anger, Disgust, Fear, Happiness, Sadness, Surprise]</td>
</tr>
<tr>
<td>Non-neutral Expression Intensities</td>
<td>4 [1 (lowest) to 4 (highest)]</td>
</tr>
<tr>
<td>Neutral Expression</td>
<td>1 [at a single intensity]</td>
</tr>
<tr>
<td>Yaw Angle Variations</td>
<td>13 [-90,-75,-60,-45,-30,-15,0,15,30,45,60,75,90]</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100 × ((6 × 4) + 1) × 13 = 32,500 Images</strong></td>
</tr>
</tbody>
</table>

**Table 3: Numerical breakdown of RIS contents**

Part 2: Verified Image Set (VIS)

Faces and facial landmarks identified from images in the RIS

The images in the RIS each contain one face displaying a known expression with a known intensity at a unique pose. However, not all faces in the RIS were detected. Furthermore, the pose generated in the RIS did not always match the yaw and roll angles detected by PittPatt.

For each detected face the following meta-data was collected: Face centre coordinates, face height, face width, face yaw angle, face roll angle, as well as several facial landmark coordinates depending on the detected yaw angle as shown in Table 4.

<table>
<thead>
<tr>
<th>Yaw Range</th>
<th>Potential Landmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>-90° &lt; Yaw &lt; -36°</td>
<td>Left Eye, Nose Bridge, Eye Nose, Left Upper Cheek, Left Lower Cheek</td>
</tr>
<tr>
<td>-36° ≤ Yaw ≤ 36°</td>
<td>Right Eye, Nose Base, Eye Nose, Right Upper Cheek, Right Lower Cheek</td>
</tr>
<tr>
<td>36° &lt; Yaw &lt; 90°</td>
<td>Right Eye, Nose Bridge, Eye Nose, Right Upper Cheek, Right Lower Cheek</td>
</tr>
</tbody>
</table>

**Table 4: Potential landmarks extracted by PittPatt in each of 3 yaw ranges (adapted from [1])**

Following PittPatt’s detection process several operations were required to produce a consistent training set. The first of these operations involved filtering out incorrectly detected or entirely
undetected faces. Overall, a total of 3,829 images were discarded after faces with low detection confidence were filtered manually.

The second operation involved 'binning' the detected faces according to the yaw value chosen by PittPatt. In order to make use of the potential landmark set in each yaw range shown in Table 4, while maintaining consistency with the 13 discrete views generated for the RIS, 12 yaw bins were set up with ranges as shown in Figure 10.

Finally, since the 3D VRML models contained in the BU-3DFE database are not symmetrical about the facial midline, all images were flipped horizontally and added to the database with their adjusted metadata. A histogram of the number of images contained in each yaw bin before and after this process is shown in Figure 11.
The asymmetry in the 3D models is evident in the inequality in image counts in opposing bins (1, 12), (2, 11), (3, 10), (4, 9), (5, 8), and (6, 7). The symmetry of the face detector allows for the horizontal flipping of the detected faces from each yaw bin and including the flipped version in the opposing bin. In addition, the performance of the face detector is clearly stronger around a frontal pose and tapers off as the yaw angle increases in magnitude.

Overall, the VIS contains 57,342 images as per the expression/intensity breakdown shown in Table 5.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Intensity</th>
<th>Per Intensity</th>
<th>Per Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>1</td>
<td>1,157</td>
<td>4,688</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1,195</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1,168</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1,168</td>
<td></td>
</tr>
<tr>
<td>Disgust</td>
<td>1</td>
<td>1,170</td>
<td>4,649</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1,172</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1,150</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1,157</td>
<td></td>
</tr>
<tr>
<td>Fear</td>
<td>1</td>
<td>1,164</td>
<td>4,557</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1,151</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1,125</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1,117</td>
<td></td>
</tr>
<tr>
<td>Happiness</td>
<td>1</td>
<td>1,170</td>
<td>4,548</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1,163</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1,106</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1,109</td>
<td></td>
</tr>
<tr>
<td>Sadness</td>
<td>1</td>
<td>1,155</td>
<td>4,642</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1,174</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1,167</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1,146</td>
<td></td>
</tr>
<tr>
<td>Surprise</td>
<td>1</td>
<td>1,135</td>
<td>4,452</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1,137</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1,111</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1,069</td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>1</td>
<td>1,135</td>
<td></td>
</tr>
<tr>
<td>Total Before Flipping</td>
<td>28,671 Images</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total After Flipping</td>
<td>57,342 Images</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Numerical breakdown of the VIS contents according to expression and intensity
Part 3: Cropped and Resized Image Set (CRIS)
Faces manipulated according to facial landmark statistics

An important characteristic of a useful facial database is spatial consistency across images. Human facial features vary between individuals but the general structure (or ordering) of features is quite standardised, as noted in ‘Result 6’ of Sinha et al.’s summary of the face recognition field [135]. Warping faces to match a standard template is a popular method for creating this spatial consistency, although, simple cropping and resizing of faces produced good results using the metadata provided from PittPatt. The spatial consistency can be seen in the mean feature analysed faces presented in Appendix IV.

The images in the VIS contain tags of facial landmarks detected by PittPatt. Each detected face’s roll angle was reversed through in plane rotation, resulting in a set of face images with 0° roll. As shown in Table 4, the face’s yaw angle controls the number and type of identified landmarks. An investigation of the cropping parameters according to the geometric location of the landmarks was carried out for each yaw bin in Figure 10. The cropping parameters were loosely based upon the values provided in [136]. For each yaw bin, a sample of 50 cropped images was observed, along with the mean greyscale images calculated across the entire image set for each bin.

Table 6: Cropping parameters for the 12 yaw bins. The dashed black box denotes the face region detected by PittPatt with height $H$ and width $W$ (adapted from [1]). LEX = Left Eye x coordinate, LEY = Left eye y coordinate, RE = Right Eye, NBR = Nose Bridge, NBA = Nose Base
Following cropping, average image widths and heights were calculated for each bin, and the cropped images were resized to their respective bin’s average dimensions. The final relative crop window coordinates are shown in Table 6 and the average bin dimensions are shown in Table 7.

<table>
<thead>
<tr>
<th>Bins</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>126</td>
<td>117</td>
<td>119</td>
<td>124</td>
<td>127</td>
<td>133</td>
<td>133</td>
<td>127</td>
<td>124</td>
<td>119</td>
<td>117</td>
<td>126</td>
</tr>
<tr>
<td>Width</td>
<td>44</td>
<td>43</td>
<td>43</td>
<td>47</td>
<td>59</td>
<td>76</td>
<td>76</td>
<td>59</td>
<td>47</td>
<td>43</td>
<td>43</td>
<td>44</td>
</tr>
</tbody>
</table>

Table 7: Face resize dimensions for the 12 yaw bins (all measurements are in pixels)

It is important to note that PittPatt may not detect all the facial landmarks shown in Table 4. Also, as shown in Table 6, the cropping process did not involve all potential facial landmarks. Thus, images in the VIS which did not have the required landmarks labelled were discarded from consideration for the CRIS. The CRIS contains a total of 50,158 images, meaning 7,184 images were discarded due to lack of landmarks. A sample image after the crop and resize process is shown in Figure 12.

![Sample cropping process for a face in the VIS. Original image (left), cropped image (centre), final cropped and resized image (right)](image)

Figure 12: Sample cropping process for a face in the VIS. Original image (left), cropped image (centre), final cropped and resized image (right)

VII-2 Feature Analysis and Dimensionality Reduction

Feature Analysis

Prior to any feature analysis taking place, the input image (training or testing) was histogram equalised in an effort to remove illumination variation effects. This process involved brightness normalisation across the full image while increasing contrast between different image regions.
Subsequently, Gaussian smoothing using a $5 \times 5$ kernel was carried out to remove discontinuities in the image.

Feature analysis was carried out using one of the three operators shown in Table 8. In the case of Sobel, the image was convolved with each directional operator, and the two outputs were then summed, while for Discrete Laplace, the image was simply convolved with the operator. Logical comparison of neighbourhood pixel values was carried out for Local Binary Pattern Analysis as shown in Table 8. The output of the analysis stage produces a feature image, which is then flattened into a one dimensional feature vector.

<table>
<thead>
<tr>
<th>Sobel (S)</th>
<th>Discrete Laplace (DL)</th>
<th>Local Binary Patten (LBP)</th>
</tr>
</thead>
</table>
| $S_x = \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 2 & 8 & 12 & 8 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ -2 & -8 & 12 & -8 & -2 \\ -1 & -4 & -6 & -4 & -1 \end{bmatrix}$ | $\begin{bmatrix} -1 & -3 & -4 & -3 & -1 \\ -3 & 0 & 6 & 0 & -3 \\ -4 & 6 & 20 & 6 & -4 \\ -3 & 0 & 6 & 0 & -3 \\ -1 & -3 & -4 & -3 & -1 \end{bmatrix}$ | $\begin{bmatrix} p_0 \ p_1 \ p_2 \\ p_7 \ p_c \ p_3 \\ p_6 \ p_5 \ p_4 \end{bmatrix}$ for each pixel $c$
| $S_y = \begin{bmatrix} -1 & -2 & 0 & 2 & 1 \\ -4 & -8 & 0 & 8 & 4 \\ -6 & -12 & 0 & 12 & 6 \\ -4 & -8 & 0 & 8 & 4 \\ -1 & -2 & 0 & 2 & 1 \end{bmatrix}$ | $\begin{bmatrix} -3 & 0 & 6 & 0 & -3 \end{bmatrix}$ if $p_c \geq p_i$ |
| $\text{LBPC} = \text{LBPC} + 2^i \end{bmatrix}$ for $i = 1:8$ end |

Table 8: Feature Analysis Operators

Appendix III shows a sample image from the CRIS, its original full-size counterpart from the BU-3DFE database and their respective feature images. Note that these feature images were normalised to 256 grey levels strictly for viewing, while the actual feature images used for processing were left un-normalised.

Appendix IV shows the mean expression-specific image for each yaw bin under the different features. The overall mean image for each feature in yaw bin 6 is shown in Table 9 on p. 54.

Principal Component Analysis (PCA)

Prior to training the classifier, the CRIS images were analysed for features, and the feature vectors stacked into large feature matrices, one for each yaw bin. PCA was then carried out on each matrix individually to uncover the most discriminant features, capturing the largest amount

---

10 Other feature operators investigated include Circular Local Binary Pattern, $3 \times 3$ Sobel Operator, $3 \times 3$ Discrete Laplace Operator, Gabor Real, Gabor Imaginary and Gabor Magnitude Filters
of variance features. These reduced feature matrices constituted the *feature analysed and reduced* CRIS (FAR-CRIS). A more detailed explanation of the benefits of using PCA is given in [138].

Projection of the feature matrix in each bin onto its respective PCA generated eigenspace resulted in a *reduced feature matrix*. The first 100 eigenvalues and eigenvectors were chosen for the projection process. This number allow for a high number of variables to be randomly selected for consideration during node splitting in RF construction. Additionally, an eigenspace of dimension 100 resulted in low reconstruction error, as shown in Figure 13, while requiring negligible processing time (see Section IX-5). The first 5 eigenfaces$^{11}$ for each feature in yaw bin 6 are shown in Table 9.

![Figure 13: Per pixel greyscale intensity reconstruction error for varying number of components](image)

Figure 13: Per pixel *greyscale intensity* reconstruction error for varying number of components

---

$^{11}$ Numbered in order of decreasing variance coverage
Table 9: Mean feature Image and first 5 eigenfaces for each feature in yaw bin 6
VII-3 Training the Classifier

As mentioned in the methodology summary, the structure used to determine the expression displayed on a face in an image consists of two classifier stages. The first stage involves a collection of binary RFs based on yaw bins, where each possible expression (or expression-intensity)–yaw bin combination is represented using an individual RF. Thus, depending on the type of forest collection, under the binary scheme each individual forest is geared to assert or deny the presence of an expression or expression-intensity combination.

Once a feature vector representing an image has been evaluated by the RF collection, a posterior probability vector is obtained. This vector is the input to an SVM (or k-NN) labeller that produces a final descriptor for the face.

Evaluator - Random Forests (RFs)\textsuperscript{12}

As a classifier, RFs rely on large scale voting by groups of decision trees to estimate the posterior probability density function for a given observation vector. The core strength of the algorithm lies in the speed at which a high number of trees can be evaluated. A series of simple threshold comparisons at the decision nodes of each tree determines the leaf node for the input feature vector. A local posterior distribution determined during the training stage is associated with each leaf node of each tree in the RF. This tree-specific local posterior distribution pertains to the expressions or expression-intensity pairs for the given feature vector. A forest-specific posterior distribution is obtained by combining the tree-specific local distributions across the entire RF.

Construction of each tree in the forest involves random selection of training vectors from the full training set, followed by random selection of features to consider for splitting at each decision node. During training, the features and corresponding threshold values picked to ultimately direct samples at each node maximise an information gain measure. This measure is evaluates how well a particular choice of feature-threshold pair can separate different training vectors of different classes. Furthermore, the measure used must accommodate the feature’s type (continuous or discrete) and range. Information gain measures rely on entropy minimisation following node splitting. However, in cases where features can take on a large range of values, the effectiveness of the pure information gain measure is diminished. Alternatively, an information gain ratio is at times employed to counteract the effects of large ranges of feature values.

\textsuperscript{12} An extended summary of Random Forests is provided in Appendix XI.
The OpenCV 2.3 framework [139] contains an implementation of RFs called ‘CVRTrees’. Our program used the Gini impurity measure, which is the default node split evaluator in this template. In the offline training process employed in this thesis, the best split at each node was limited to 50 features selected at random for each node. Since the Eigenspace method used for reducing the feature matrix was set to a dimension of 100, half the features were ignored at each split. Additional parameters used in forest construction are shown in Table 10.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of features to consider for splitting</td>
<td>50</td>
</tr>
<tr>
<td>Maximum tree depth</td>
<td>20</td>
</tr>
<tr>
<td>Minimum number of samples per node to split on</td>
<td>5</td>
</tr>
<tr>
<td>Minimum forest accuracy</td>
<td>99 %</td>
</tr>
<tr>
<td>Number of random samples per tree</td>
<td>⅓ of full set</td>
</tr>
<tr>
<td>Number of trees per forest</td>
<td>100, 200 and 300</td>
</tr>
</tbody>
</table>

Table 10: Parameters used for RF construction

The minimum forest accuracy parameter is computed by evaluating each tree on the out of bag samples during the training stage. If this minimum accuracy is not achieved, the forest is discarded and a new one is constructed. The number of trees per forest was varied to be 100, 200, or 300. The ultimate number of trees per forest was chosen according to the performance of the full classifier consisting of the RF collection and labellers (see p. 58).

The RFs were trained with a bias towards avoiding false positive classifications by providing the negative samples with a higher weight during training. The weight ratio of negative to positive samples was set to 6:1. This was achieved following experimentation, which indicated that much smaller (e.g., 1:4) or much larger (e.g., 25:1) ratios produced significantly worse accuracy rates.

The parameters in Table 10 were used to produce RF collections, for each yaw bin, under three different training arrangements.

1. Each unique expression-intensity subset was again pitted against all other training vectors as but not including the other intensities of the same expression, in a binary classification scheme for each forest. This model does not attempt to assert a relation between different intensities of the same expression, but does attempt to assert inter-expression independence.

2. Each unique expression-intensity subset was pitted against all other training vectors (including other intensities of the same expression). This model assumes intra-expression independence with respect to intensity, in addition to inter-expression independence for Ekman’s six.
3. All intensities of a certain expression were pitted against all other expressions and their intensities. This model assumes intra-expression non-independence with respect to intensity.

For RF collections of types 1 and 2, the forest count was 25 RFs per yaw bin, or 300 RFs (12 yaw bins × 25 expression-intensity combinations\(^{13}\)) in total. As for type 3 collections, each yaw bin contained 7 RFs, totalling 84 RFs (12 yaw bins × 7 expressions) across all bins.

The output of each forest in each collection for the relevant yaw bin is a fraction quantifying the chances that the input vector displays the respective expression (type 3) or expression-intensity combination (types 1 and 2) for the forest that was trained. This value represents a forest specific posterior probability.

Concatenating the output of all forests of any one of the three collections in the relevant yaw bin produces a vector of forest specific posterior probabilities. Normalising the vector by dividing each element by the vector’s 1-norm produces a posterior probability distribution function (poPDF) for the collection.

For the first and second type of RF collections, the output is a 25 element poPDF such that a probability value is obtained for each unique expression-intensity combination. The output of the third type of RF collection results is a 7 element poPDF containing a value for each of the six expressions.

Labeller – Support Vector Machine (SVM) or k-Nearest Neighbours (k-NN)

One way to determine a final label would be to select the expression corresponding to the maximum value from the poPDF vector. However, experimentation with the poPDF elements produced noticeably better performance on a test subset of the FAR-CRIS. Thus, another classification tool, which acted as a labeller, was employed to circumvent the need for manual scaling of the poPDF.

All the RFs were trained on the full FAR-CRIS, with the required forest construction accuracy maintained at the level shown in Table 10. The full FAR-CRIS was then evaluated by the RFs and the resultant poPDF vectors (CRIS-poPDF) obtained from each forest were used to train the subsequent SVM or k-NN labeller.

\(^{13}\) See Table 5
The output of type 3 RF collections was used to obtain a single expression label (1 of 7 class labeller). The output of type 1 and 2 RF collections was employed under 3 schemes:

1. The 25 poPDF values were used to produce a 1 of 25 class labeller resulting in expression-intensity classification.
2. The 25 poPDF values were used to produce a 1 of 25 class labeller. The label was then stripped of its intensity value, ultimately resulting in a simple 1 of 7 expression classification.\(^\text{14}\)
3. The 25 poPDF values were used to produce a 1 of 7 class labeller, selecting an expression without specifying the intensity directly.

The difference between the second and third labelling schemes is the true training labels used to train the SVM or k-NN in each case. For the second scheme, much like the first, the training labels contain both expression and intensity information, while the training labels of the third scheme contained only expression information.

In the case of the SVM, the kernel used was a radial basis function under a C-support scheme. The parameters of the SVM were obtained via grid optimisation, carried out through the ‘CvSVM’ OpenCV [139] class. The SVM was trained on half the CRIS-poPDF vectors for each bin, selected at random while ensuring class coverage. The k-NN was implemented using the ‘CvKNearest’ OpenCV [139] class where the label selected was determined using simple Euclidean distance to the training samples.

Selecting the Classifier Parameters

As mentioned earlier in this section, the optimal classifier parameters for both RF and labeller were chosen according to performance over the test half of the CRIS-poPDF. The criterion used to determine performance was the classification accuracy over this test set. The three tree counts (100, 200 and 300) were tested for the three collection types (p. 56) under the three labeller training schemes (p. 58).

\(^\text{14}\) This was not used with the k-NN labeller, due to redundancy.
Figure 14 through Figure 22 show the results of SVM labelling for the various RF collection types and labelling schemes under Sobel feature analysis. Note that the same CRIS-poPDF split was used to test the performance of the Sobel feature k-NN labeller, as shown in Figure 33 through Figure 41 in Appendix V. As mentioned earlier, the second labelling scheme was discarded for the k-NN labeller since it produced accuracy results similar to those obtained under the third labelling scheme.

Figure 14: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 100 trees per forest under the three different labelling schemes using Sobel features

Figure 15: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 100 trees per forest under the three different labelling schemes using Sobel features

Figure 16: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 100 trees per forest using Sobel features
Figure 17: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 200 trees per forest under the three different labelling schemes using Sobel features.

Figure 18: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 200 trees per forest under the three different labelling schemes using Sobel features.

Figure 19: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 200 trees per forest using Sobel features.
Figure 20: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 300 trees per forest under the three different labelling schemes using Sobel features.

Figure 21: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 300 trees per forest under the three different labelling schemes using Sobel features.

Figure 22: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 300 trees per forest using Sobel features.
The criterion used to select the best classifier structure was the highest average accuracy calculated across the four 'central' yaw bins (5, 6, 7 and 8). The reduced sample size of the off-centre yaw bins is thought to be a partial reason for the generally parabolic accuracy curves obtained from parameter tuning. However, exhaustive testing is required to ascertain the cause of this reduction in correct classification. We note that this type of fall-off has also been observed in the current literature for face recognition. The accuracies for the Sobel feature classifiers with SVM and k-NN labellers are shown in Table 11 and Table 12, respectively. Note that all accuracies in the tables were obtained using only the selected testing portion of the BU-3DFE database.

<table>
<thead>
<tr>
<th>Forest Type</th>
<th>Number of Trees</th>
<th>Labelling Scheme 1</th>
<th>Labelling Scheme 2</th>
<th>Labelling Scheme 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>50.3360</td>
<td>75.4200</td>
<td>76.7766</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>54.6142</td>
<td>77.6073</td>
<td>76.4437</td>
</tr>
<tr>
<td>1</td>
<td>300</td>
<td>55.6689</td>
<td>77.3958</td>
<td>77.0162</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>55.6752</td>
<td>75.2676</td>
<td>77.0193</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>58.4132</td>
<td>77.2744</td>
<td>77.3771</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>59.8102</td>
<td>77.1873</td>
<td>78.2701</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>--</td>
<td>--</td>
<td>73.2078</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>--</td>
<td>--</td>
<td>73.9172</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>--</td>
<td>--</td>
<td>74.3559</td>
</tr>
</tbody>
</table>

Table 11: Central bins (5, 6, 7 and 8) mean classification accuracies for Sobel feature Classifiers with SVM labelling (all values are %)

<table>
<thead>
<tr>
<th>Forest Type</th>
<th>Number of Trees</th>
<th>Labelling Scheme 1</th>
<th>Labelling Scheme 2</th>
<th>Labelling Scheme 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>32.8023</td>
<td>--</td>
<td>65.6855</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>35.9699</td>
<td>--</td>
<td>67.4653</td>
</tr>
<tr>
<td>1</td>
<td>300</td>
<td>36.3806</td>
<td>--</td>
<td>68.0129</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>38.6085</td>
<td>--</td>
<td>64.8391</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>42.5353</td>
<td>--</td>
<td>67.9507</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>43.5621</td>
<td>--</td>
<td>68.2432</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>--</td>
<td>--</td>
<td>63.7065</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>--</td>
<td>--</td>
<td>65.2187</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>--</td>
<td>--</td>
<td>65.2499</td>
</tr>
</tbody>
</table>

Table 12: Central bins (5, 6, 7 and 8) mean classification accuracies for Sobel feature Classifiers with k-NN labelling (all values are %)

From Table 11 and Table 12, it is clear that the k-NN labeller’s performance was significantly below that of its counterpart. Thus, the k-NN labeller was abandoned for the construction of the discrete Laplace and the LBP classifiers, as well as the testing of unconstrained image databases. In addition, due to the improved accuracy of 200 and 300 trees/RF collections over those containing only 100 trees/RF, the latter collections were also abandoned. Classification results and accuracy tables for discrete Laplace and LBP feature analysis are shown in Appendix.
VI and Appendix VII, respectively. The optimal classifier structure (parameters, collection type and labelling scheme) for Sobel, discrete Laplace and LBP is contained in Table 13.

<table>
<thead>
<tr>
<th></th>
<th>Sobel</th>
<th>Discrete Laplace</th>
<th>Local Binary Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF Collection Type (p. 56)</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Number of Trees per Forest</td>
<td>300</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>Labeller Type</td>
<td>SVM</td>
<td>SVM</td>
<td>SVM</td>
</tr>
<tr>
<td>Labelling Scheme (p. 58)</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Central Bins Accuracy</td>
<td>78.2701 %</td>
<td>76.4343 %</td>
<td>78.9172 %</td>
</tr>
</tbody>
</table>

Table 13: Optimal classifier structure for SVM based labelling

Although admittedly it is a comparison between 2D and 3D expression recognition, the accuracy rates reported for frontal poses compare favourably to those reported by the BU-3DFE authors in [140], where manual segmentation of the full 3D model of the face, followed by manual landmark extraction and feature analysis produced an 83.6% accuracy rate.

VII-4 Temporal Integration

This section outlines the method used to analyse facial expressions displayed by an individual over a sequence of images or a number of consecutive video frames. The temporal consistency of a computed expression (somewhat related to the person’s underlying mood) would be expected to be a strong indication of the consistency of the expression recognition system.

PittPatt provides a subject tracking framework that supplies different individuals in a single video with a unique identification (ID) number. Thus, by evaluating faces in a frame using the same technique used to process single images (see Flowchart 4 in Appendix II), while accumulating the result for individuals from frame to frame, a progression of each person’s facial expression can be obtained. To carry out online expression classification while benefiting from sequential frame to frame relationships, a histogram of the resultant labels in a window surrounding each frame is analysed and a single label is picked based on a popularity measure.

Figure 23 contains a graphic visualisation of the temporal integration process for sequential frames. Given that each frame expression label was obtained using the classification results from $N$ frames, an initialization procedure is required preceding histogram processing.
Biological motivation for using a temporal window to obtain single frame labels is heavily rooted in the experimental work of Haberman et al. in [141]. The authors discuss the principles of Gestalt grouping [142] and the importance of low level features, such as edges, in face representation in the human brain. Experiments by Haberman et al. have identified the duration of the presence of a single expression as one of the main variables required for successful labelling [141]. These findings support the mechanism used in this thesis, where non-sustained expression changes were smoothed using an accumulated histogram. For example, if an individual displays a happy expression for an extended period of time, frowns in anger for a single frame and then returns to being happy, the averaging process would ignore the frown since the mode of the histogram of expressions is always chosen. If no temporal integration scheme were employed, the frown would be taken into account, representing an anomaly or discontinuity in expression. Furthermore, if the individual actually continued to frown, the temporal integration scheme would gradually shift to asserting an angry expression once enough frowning frames have been accumulated.

Note that ties in the temporal integration histogram were decided by giving preference to maintaining a consistent expression label, if possible, and by random selection when the label of the previous frame was not available.

**Figure 23: Temporal Integration Scheme Visualisation**
The system was designed as a completely autonomous entity, requiring no manual intervention for classifying still images or video streams. The databases listed in Table 2 on p. 34 were tested using the procedure outlined in Flowchart 4 on p. 98 using a number of various classifier combinations, including the optimal version characterised by the parameters in Table 13. In cases where the test material was a video, the temporal integration scheme utilised the output of single image classification in the manner shown in Figure 23. The results of this testing procedure were compiled in Chapter VIII (accuracy plots are in Appendix VIII and Appendix IX), followed by a broad discussion and review of the findings of both the training and testing portions in Chapter IX.
VIII Results

The results reported in this chapter are the product of the testing strategy briefly described in Section VII-5. Some clarifications related to the overall platform used for testing are listed below. Plots of classification accuracy for the evaluation of still image databases are provided in Appendix VIII. Plots for video classification accuracy are provided in Appendix IX.

1. The full classifier was used at all times, meaning that no partitioning of the RF collections was carried out when test databases were known to contain only a subset of the 7 expressions.

2. According to available reports and the dispersed geographic locations of database collection, the probability of a subject being present in both training and testing databases is quite low.

3. Classification was carried out using all three types of 200 and 300 trees/RF collections for all features. SVM based labelling schemes 2 and 3 from p. 58 were used to extract a final label from the poPDF vectors. The complete evaluation results are available online\(^\text{15}\), while those obtained using the optimal classifiers from Table 13 are shown in this thesis.

4. Due to the very short period of displaying the intended expression in video clips as well as the lack of frame by frame labels, a correct classification was acknowledged during video testing if the first or second most popular label across the entire video, matched the true expression label. The reason for accepting the second most popular label is because it is common for laboratory captured databases to contain subjects progressing from a neutral expression, to the peak intended expression and then back to the neutral expression. Thus, in such a progression, the neutral expression would rank first, while the peak intended expression would come in second. The results of using only the most popular label are also provided online.

\(^{15}\) http://www.cim.mcgill.ca/~mmosta3/
IX Discussion

The main purpose of the project was to produce a real-time unconstrained facial expression recognition framework. By-products of this endeavour were a number of observations related to the various topics discussed in Chapters III through VII.

IX-1 Expressions and Expression Intensities

The confusion matrices shown in Figure 24 pertain to the classifier combinations shown in Table 11. Normalised greyscale intensity matrices are shown here to simplify visualisation of the results. Note that the off-diagonal values are barely visible, indicating the overall consistency of the results. The numerical values from which these matrices were generated are given in Appendix X.

Figure 24: Sobel (left), discrete Laplace (centre) and local binary pattern (right) expression confusion matrix. Note that darker shades indicate higher counts.

These confusion matrices were obtained by incrementing the cell indexed by the true expression of the image (vertical axis) and the expression label obtained by the detection framework (horizontal axis). The matrices were row-normalised, thus, reading each matrix row by row gives an indication of the relative level of confusion of each expression with its counterparts.

The greyscale levels across the three features in Figure 24 are quite similar indicating that the relative inter-expression confusion is stable from feature to feature. In the case of Sobel and discrete Laplace features, the confusion similarity is of no surprise, given that both use geometric kernels based on derivatives.
The partial visual symmetry present in all three matrices in Figure 24 indicates that a consistent relationship exists between certain expressions. Quantifying this hypothesis involved calculation of the correlation coefficient between the various detected expressions. These are shown in Figure 25.

![Correlation Coefficients Chart](chart.png)

**Figure 25: Correlation coefficients for the 7 expressions. Note: the correlation coefficient sum (Σ) does not include self-correlation.**\(^{16}\)

The inter-expression correlation was calculated for the results of the optimal classifiers from Table 13 using the values shown in Appendix X and the formula shown in Equation 1. These correlations were subsequently averaged across the three features to produce the chart shown in Figure 25.

---

\(^{16}\) AN: Anger, DI: disgust, FE: Fear, HA: Happiness, NE: Neutral, SA: Sadness, SU: Surprise
\[ R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(i,j)}} \]

Equation 1

As observed in the figure, no positive correlation coefficient was found to exist between any two different expressions, although small negative coefficient values were found for anger-sadness, anger-disgust, neutrality-sadness and fear-happiness. Each of these coefficient values suggests a negative relationship between the expression pairs when considered independently. Taking the anger-sadness coefficient for example, the following could be stated: In the presence of a true anger expression, a negative correlation with sadness exists, suggesting they are antagonistic. However, when compared to the remainder of the coefficient pool, a different relationship could be implied. Again, taking the example of anger-sadness, along with the much more negative anger-happiness coefficient, the relationship could be worded as follows: In the presence of a true anger expression, a negative correlation with sadness still exists; however, it is more positive than the correlation between happiness and true anger, suggesting that anger and sadness are less unrelated than anger and happiness. Thus, the significance of the correlation coefficients comes from their comparative relationships rather than absolute ones.

The framework implemented here renders the sum of correlation coefficients quite useful since it quantifies the ease of identifying each of the seven expressions. Figure 25 shows this sum without including the self-correlation. The more negative a correlation coefficient sum, the more independent the respective expression. Thus, it is valid to say that surprise, happiness and neutrality are most easily detected, while anger and sadness are the most difficult to identify.

Langner et al. [95] provided the figure shown in Figure 26, obtained by surveying a group of students regarding the expressions they thought subjects from the RaFD image set were displaying. The relations suggested by Figure 25 (anger-sadness, anger-disgust, neutrality-sadness and fear-happiness) do not entirely agree with those shown in Langner's work, however, anger-disgust confusion is quite prominent in both figures.
It is worthwhile to emphasise that Figure 26 was generated through human polling with an average accuracy of 81.75%. In [143], Goeleven et al. studied expression classification by humans and reported an accuracy of 71.75% with a distribution similar to that obtained using the framework developed for this thesis, as well as that obtained by Langner et al. in [95]. The implication of the above discussion is that single still images of subjects displaying facial expressions may not be sufficient for proper classification of the underlying emotions. In reality, the key for humans to properly assess facial expressions around them is temporal context, which cannot be extracted from a single image. Temporal context is revisited in Section IX-3 and in Chapter X.

Nevertheless, Langner’s experiments may provide an indication as to the importance of a large sample base for training, which is a suspected cause of low classification accuracy rates in the off-centre yaw bins, as shown in Figure 11. The high classification rates obtained by humans, even in off-centre yaw bins seem to support this suspicion since the training sample set for humans is extensive and encountered throughout a lifetime.
The first two RF collections listed in Section VII-3 paired with the first labelling scheme shown in the same section, were structured in a manner aimed at uncovering the validity and subsequent significance of the importance of expression intensity, and how intensities affect one another.

A preliminary indicator of the existence of a relationship between expression intensities is the discrepancy in classification accuracy between the first two types of RF collections. Table 11 shows that the classification accuracy increases by a little over 4% when using type 2 RF collections instead of type 1. The improvement in accuracy is represented by lighter off-diagonal greyscale intensities in the confusion matrices in Figure 27. Lighter off-diagonal greyscale intensities mean that more samples were correctly classified, and were accounted for in the diagonal elements.

![Figure 27: Confusion matrix for type 1 RF collections (left) and type 2 RF collections (right) with 300 trees/RF using Sobel features and type 1 labelling scheme](image)

However, for both collections in Figure 27 it is clear that the bulk of the misclassification occurs between intensities of the same expression as visually represented by expression boxes lying along the diagonal. The boxes are present due to the arrangement of the expression labels along the two axes. The information gained from this accuracy discrepancy is twofold. Firstly, intensities of the same expression possess differentiating characteristics evident in the improvement in accuracy from type 1 to type 2 RF collections. Secondly, intensities of the same expression share a number of characteristics evident in the continued presence of the expression boxes in the confusion matrix on the right in Figure 27. The second point is better visualised in the binarised correlation coefficient plot shown in Figure 28.

71
Figure 28: Binary correlation coefficients for the 25 intensities calculated from the confusion matrix on the right in Figure 27. Black represents positive correlation, white represents negative correlation.

Figure 28 illustrates that it can be expected that various intensities of the same expression will be confused with each other during classification. In addition, Figure 27 and Figure 28 give theoretical credence to labelling scheme 2 in Section VII-3.

The off-diagonal positive correlation coefficients in Figure 28 can be explained through the lowest intensity frontal images of anger, disgust, sadness and neutrality, shown in Figure 29. The lack of strong visually discriminative features between the four sample images speaks to the proximity of these low intensity expressions. Similar arguments could be made for low intensities of fear and happiness, as well as stronger intensities of anger and disgust.

In the case of confusion between fear and happiness, it is not uncommon to see smiles on the lips of the subjects as a result of nervousness, which is a well documented human reaction [144] and can be seen on the right in Figure 29. Clear examples of scared individuals smiling can be seen in the FEED video database set [99].

Figure 29: Lowest intensity face image from BU-3DFE database for [anger, disgust, neutral, and sadness] and [fear and happiness] displayed by subject ‘M0014’
Finally, considering the plots in Figure 30 extracted from the confusion matrix on the right in Figure 27, it is clear that the relationship between intensities diminishes as the distance between them increases. For example, the strength of the relation (represented by greyscale intensity) between AN1 and AN2 is higher than the relation between AN1 and AN4. These matrices confirm the postulate that expressions obtained from images labelled by both humans and computers are graded along some scale, which could be used to construct a continuous dimensional expression space. This space would have its origin at a neutral expression, anchoring six axes, one for each expression. Traversal along an expression axis away from the origin would represent an increase in expression intensity.

Figure 30: Auto-confusion matrices for anger, disgust and happiness taken from the matrix on the right in Figure 27
Still Image Analysis

The contents of databases used for testing varied greatly in image size, image quality, camera characteristics, lighting, background, subject demographics, as well as interpretation and demonstration of the various expressions. Uncovering the exact influence each of these variables had on the accuracy of the unconstrained classification system is extremely difficult. Such an undertaking is beyond the scope of this thesis. Instead, the discussion here focuses on a number of high-level observations gained from inspection of the accuracy charts in Appendix VIII.

Firstly, we observed during training of the classifier, that the central yaw bins consistently experienced higher accuracies than other yaws. This is particularly obvious from the results for the database that were created under similar conditions\textsuperscript{17}, KDEF [94] and RaFD [95].

Furthermore, as the human ratings in Figure 26, as well as the correlation coefficient sums in Figure 25 implied, happiness and surprise were the most identifiable expressions in still images. Whether that speaks to their ‘universality’ in the Ekman sense, as opposed to being less identifiable as expressions (such as fear and sadness) is an interesting question.

The results for disgust were a departure from the general trend of human classification. However, when considering the correlation coefficient sums, which indicated that disgust is neither the most difficult nor the easiest to identify, the results are not as anomalous as Figure 25 may indicate.

Table 14 shows the highest accuracies obtained for the central yaw bins (5, 6, 7, and 8), across all the classifier combinations. The logic for using central yaw bin accuracy is the same as that employed for classifier evaluation during training (see Section VII-3). The results in this table indicate that the level of difficulty shown here is different from the earlier situation in which the same databases were used for both training and testing. We observe a significant drop in the ‘best accuracy’ in the first row for the test databases in Table 14, compared to those obtained for the BU-3DFE database shown in Table 11. The reasons for this discrepancy could be that our classification framework was unable to generalise sufficiently to correctly classify more of the test databases. The second reason for this discrepancy could be the inconsistency of the interpretations made by the human labellers. For example, an angry expression in a test database may actually be closer to a disgust expression in the BU-3DFE. Clearly, these reasons are intrinsic to the experimental model currently being employed in the literature and in this

\textsuperscript{17} Similar construction relates to regular variation in pose and expression under consistent lighting and background conditions.
thesis. However, given that the accuracies were all above chance \(1/7 \approx 14\%\), some significantly so, the indication is that the framework presented in this thesis does achieve a degree of success.

<table>
<thead>
<tr>
<th>Database</th>
<th>KDEF</th>
<th>RaFD</th>
<th>JAFFE</th>
<th>PICS – pain</th>
<th>SFEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Central Bins Accuracy</td>
<td>46.17 %</td>
<td>55.03 %</td>
<td>41.96 %</td>
<td>35.81 %</td>
<td>20.57 %</td>
</tr>
<tr>
<td>Feature</td>
<td>Sobel</td>
<td>Sobel</td>
<td>Sobel</td>
<td>Sobel</td>
<td>Sobel</td>
</tr>
<tr>
<td>RF Collection Type (p. 56)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Number of Trees per Forest</td>
<td>200</td>
<td>300</td>
<td>200</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>Labelling Scheme (p. 58)</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 14: Best central bins (5, 6, 7, and 8) accuracy obtained for each of the still image test databases along with the classifier parameters giving rise to this accuracy

A comparison of the results using the still image test database in this thesis with those from other publications was not possible. This was due to the unavailability of an appropriate platform. To the best of the author’s knowledge, the existing literature has not reported results for the databases listed in Table 14, for the situation where the training and test sets were not from the same database. Instead, the leave-one-out test strategy was employed. Nevertheless, the results from other methods are provided for comparison purposes in Table 15. Unless otherwise indicated, the methods shown in the table are unconstrained and perform in real-time.

<table>
<thead>
<tr>
<th>Database</th>
<th>KDEF</th>
<th>RaFD</th>
<th>JAFFE</th>
<th>SFEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Accuracy</td>
<td>89%</td>
<td>89.54 %</td>
<td>85.72 %</td>
<td>18.7 %</td>
</tr>
<tr>
<td>Clarification</td>
<td>Used in training Not real-time</td>
<td>Used in training Proprietary system</td>
<td>Used in training Not real-time</td>
<td>Used in training</td>
</tr>
<tr>
<td>Reference</td>
<td>[145]</td>
<td>[146]</td>
<td>[147]</td>
<td>[98]</td>
</tr>
</tbody>
</table>

Table 15: Classification accuracy achieved by other methods

In order to obtain the individual classification accuracies shown in Table 15, it was necessary to independently tune the classifier used for each database. The implication seems to be that each of these respective unconstrained databases is in some way intrinsically different from the others. This may seem to be a contradiction to the hypothesis that using an unconstrained database for training should provide some "universality", in the sense that it should produce a single robust classifier. The most likely reason that the parameters of the recognition classifier needed to be changed for each database is probably that a much larger database of unconstrained data would be required. This is beyond the scope of the thesis and is a subject for further study.

A highly pertinent example regarding the difficulty of achieving high classification rates for databases not involved in the training comes from [148]. These authors developed a framework for expression classification and trained it on two databases separately. Each database was
subsequently tested on the two different versions of the framework. Their results are shown in Table 16 and are compatible with those presented in this thesis.

<table>
<thead>
<tr>
<th>Framework Characteristics</th>
<th>Trained</th>
<th>Tested</th>
<th>Classification Accuracy Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAFFE</td>
<td>JAFFE</td>
<td>90.17 %</td>
<td></td>
</tr>
<tr>
<td>JAFFE</td>
<td>CK</td>
<td>54.05 %</td>
<td></td>
</tr>
<tr>
<td>CK</td>
<td>CK</td>
<td>90.26 %</td>
<td></td>
</tr>
<tr>
<td>CK</td>
<td>JAFFE</td>
<td>55.87 %</td>
<td></td>
</tr>
</tbody>
</table>

**Table 16: Classification accuracy of the different versions of the framework, compiled from [148]**

The significant difference in classification accuracy we obtained for the various databases was somewhat expected based on a visual examination of the images in the respective databases. Those classification results that were obtained from similar image databases\(^{18}\) to the BU-3DFE, such as the KDEF and RaFD sets, outperformed those obtained from different (PICS - pain, JAFFE) and drastically different (SFEW) databases.

As mentioned in Chapter VIII, the full classification accuracy charts are available online\(^{19}\).  

\(^{18}\) See the first paragraph of this section.

\(^{19}\) http://www.cim.mcgill.ca/~mmosta3/
The result charts for video analysis are given in Appendix IX. Fluctuations in accuracy rates from database to database, as well as from feature to feature were quite prevalent for the two labelling schemes examined (schemes 2 and 3). This may indicate that the data are inseparable under either of the labelling schemes and more investigation into the nature of the poPDF vectors is necessary.

The classification framework performed reasonably well when the BU-3DFE still-image database was used for training and the BU-4DFE video database for testing, as shown in Table 18 further below. Clearly, both originated in the same laboratory, implying that the interpretation of the expressions was probably similar. An interesting observation was the consistent misclassification of the neutral expressions with began and ended each video, thereby sandwiching the intended Ekman expression. No clear reason could be found for this problem aside from the confusion caused by low-intensity expressions and mentioned in Section IX-1. This was not an impediment to successful classification due to the overall video temporal integration and labelling method used, which acknowledged a correct classification if the second most popular detected label matched the true video label.

The FEED database [99] presents a multi-faceted challenge to the classification scheme. At times, the intended expression was labelled for an extremely short period of time, resulting in incorrect frame labelling due to the temporal integration scheme. This also produced an incorrect label for the whole video clip. In addition, we did not verify the database labels to ensure that a subject was actually displaying the intended expression. This verification step was left out on purpose in order to mimic a real world scenario. Finally, the majority of the videos in this database did not contain high intensity expressions save for disgust and happiness. As shown in Figure 29, and discussed in Section IX-1, it is quite difficult to differentiate low intensity expressions from one another.

Table 17: Video databases evaluated in this thesis

<table>
<thead>
<tr>
<th>Database, Year Published</th>
<th>Spontaneous</th>
<th>Professional Actors</th>
<th>Constrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>BU-4DFE, 2008 [90]</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>FEED, 2006 [99]</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>AFEW, 2011 [100]</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

20 Naturally, if the data are inseparable, i.e., cannot be uniquely represented in any of the feature spaces examined here, then a different feature space should be examined.
The spontaneity of the FEED database (as opposed to the relatively measured expressions of the BU-4DFE database), along with the results obtained here provide a window into true human expression display mechanisms and their misrepresentation by actors or amateurs, as is done in the majority of available databases to date.

As discussed in Section V-9, the AFEW database, consisting of unconstrained clips from feature films, presents the ultimate test environment for the classifier. Professional actors are seen as great purveyors of emotions, and their facial expressions are quite representative of these emotions. However, one would expect that accuracy would drop from the FEED image set to the AFEW image set as the subjects were predominantly filmed from off-centre yaw angles, under highly variable lighting conditions and in many cases, while reciting their lines.

The classification accuracy for the BU-4DFE was higher than for the other two databases. The general trend in accuracy levels amongst expressions was maintained, such that sadness, surprise and neutrality were still quite low, while anger and happiness remained high. The major difference in accuracy was in the classification of fear, where lower rates were found for BU-4DFE compared to its counterparts.

Table 18 provides a summary of the best accuracies obtained for each expression in the video test databases. These were selected from the full set of results and illustrate the ability of different expression classifiers to identify certain expressions under the specific database’s conditions. These results suggest that a practical application may require an initialisation step to optimise results in a real world situation. The initialisation step could take into account the most probable expressions that will be displayed, biasing the classifier towards detecting them. In addition, the image processing stage, particularly, brightness and contrast normalisation, could be geared towards the environment in which the framework is used.

<table>
<thead>
<tr>
<th></th>
<th>BU-4DFE</th>
<th>FEED</th>
<th>AFEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>96.04</td>
<td>68.52</td>
<td>56.25</td>
</tr>
<tr>
<td>Disgust</td>
<td>88.12</td>
<td>81.48</td>
<td>65.08</td>
</tr>
<tr>
<td>Fear</td>
<td>39.60</td>
<td>64.81</td>
<td>56.92</td>
</tr>
<tr>
<td>Happiness</td>
<td>70.30</td>
<td>61.11</td>
<td>40.00</td>
</tr>
<tr>
<td>Neutral</td>
<td>--</td>
<td>11.32</td>
<td>28.57</td>
</tr>
<tr>
<td>Sadness</td>
<td>65.35</td>
<td>62.96</td>
<td>29.33</td>
</tr>
<tr>
<td>Surprise</td>
<td>79.21</td>
<td>25.45</td>
<td>35.59</td>
</tr>
</tbody>
</table>

Table 18: Classification accuracy achieved by other methods (all values are in %)
The classifier presented in this thesis encountered difficulty when classifying certain video clips (particularly ‘neutral’, ‘sadness’ and ‘surprise’). The classifier’s inability to parlay successful classification of these expressions from still images to videos raises questions regarding the applicability of still image training to video evaluation. See Chapter X for a more detailed discussion of the potential advantages of a video training database and framework.

Comparison of our still image trained system with one trained on videos is provided below, albeit one trained and tested on the BU-4DFE database. Results for classification of the BU-4DFE database by the 2D method in [149] under a reduced 4 expression scheme consisting of anger, happiness, neutrality and surprise, are shown in Table 19. The results are quite similar to those we obtained from our evaluation of the test portion of the BU-3DFE still image database. Thus, although somewhat premature, the results indicate that a system trained and tested on videos, would outperform a system trained on still images and tested on videos.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>67.07 %</td>
</tr>
<tr>
<td>Happiness</td>
<td>76.27 %</td>
</tr>
<tr>
<td>Surprise</td>
<td>76.77 %</td>
</tr>
</tbody>
</table>

Table 19: Results of classifying the BU-4DFE by the 2D method in [149]

The 3D facial expression classification method described in [150] again uses a reduced 3 expression scheme consisting of happiness, surprise and sadness and was trained on samples from the BU-4DFE database gave the results shown in Table 20.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happiness</td>
<td>95.00 %</td>
</tr>
<tr>
<td>Sadness</td>
<td>91.67 %</td>
</tr>
<tr>
<td>Surprise</td>
<td>90.00 %</td>
</tr>
</tbody>
</table>

Table 20: Results of classifying the BU-4DFE by the 3D method in [150]
IX-4 Temporal Integration Scheme Interface

The temporal integration scheme, shown in Figure 23, can be seen in action for sample videos from each of the test databases online\(^\text{21}\). A short description of the strategy behind the temporal integration scheme is given below and references Figure 31.

The top left corner of the interface plays the video being analysed, with a superimposed rectangle around the face’s location in the image as detected by PittPatt [1]. Below this rectangle is an indicator of the current frame’s expression label as determined by the temporal integration scheme.

The temporal integration histogram in the bottom left corner of the interface contains the accumulated labels from a window of frames surrounding the current frame as described in Section Error! Reference source not found.. A total of five frames either side of the current frame were taken into account, bringing the total size of the window to eleven frames. The direct result of the SVM labeller for each frame in the window is added to the histogram and the mode expression is chosen as the final label for the current frame.

The charts along the middle of the interface contain the normalised poPDF vector for the current frame. Here, the outputs of the RFs are divided by their total sum, giving the user an idea of how the RF collection is classifying the current expression. The vector is split up into its respective expression and intensity labels.

The top right histogram contains an accumulation of the temporal integration scheme labels for each frame as the video progresses. The proportion of this histogram that each expression occupies as the video progresses is represented in the bottom right chart.

\(^{21}\) http://www.cim.mcgill.ca/~mmosta3/
IX-5 Computer Timing

The characteristics of the computer system used for training, testing and data storage are given in Table 21 below.

<table>
<thead>
<tr>
<th>HARDWARE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>8 x Intel Core i7 CPU</td>
<td>860 @ 2.80 GHz</td>
</tr>
<tr>
<td>Random Access Memory</td>
<td>4 x 4 GB DDR3 1333 MHz</td>
<td></td>
</tr>
<tr>
<td>Data Storage Magnetic Hard Disk</td>
<td>WD1001FAES 1TB 7200 RPM</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SOFTWARE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating System</td>
<td>Ubuntu 10.04 (Lucid) 64-bit</td>
<td></td>
</tr>
<tr>
<td>Integrated Development Environment</td>
<td>Netbeans C++</td>
<td></td>
</tr>
<tr>
<td>Compilers</td>
<td>GNU Compiler Collection (GCC) 4.4.3</td>
<td></td>
</tr>
<tr>
<td>External Libraries Used</td>
<td>PittPatt 5.0 64-bit, OpenCV 2.3 64-bit</td>
<td></td>
</tr>
</tbody>
</table>

Table 21: Characteristics of the computer system used for training, testing and storage of the classifier

A timing breakdown of the various stages of the classification process is given in Figure 32. The bulk of time required to obtain an expression label is dedicated to detecting the faces in the image along with their respective metadata. Given that the version of PittPatt face detection and tracking system used here does not employ threading, adding a more parallelised detector could improve the performance significantly. Nevertheless, the performance (about 9 fps) of the system as it stands is in real-time.

![Figure 32: Worst case scenario timing diagram](image-url)
Conclusion

This thesis has shown that the theoretical foundation upon which our framework was constructed does hold its own in real world situations. This claim is evidenced by the classification accuracies of both the BU-3DFE database, as well as those not involved in training the classifier. The accuracies were found to be competitive with the state-of-the-art methods when a platform for comparison was available. Furthermore, our framework achieved the requirements of real-time performance in an unconstrained environment, which was previously unavailable in the community.

Our biggest qualms with respect to the theoretical foundation were concerned with Ekman’s ‘universality’ of expressions. These qualms presented themselves at two distinct points as the thesis was formulated. The first point was the interpretation of expressions by posed database participants, and how the expressions relate to their spontaneous, real world counterparts.

The second point is highlighted in the realm of classification of expressions in videos. Under a still image training scheme such as ours, successful video classification is highly dependent on the assumption that the progression from one ‘universal’ expression (the source) to another (the sink) involves a sequence of intermediate expressions pertaining to the former or the latter. In reality, these intermediate expressions may contain elements of ‘non-universal’ expressions, not captured by the Ekman model. A video classification system should intrinsically accommodate these intermediate expressions through training on progressions to and from apical ‘universal’ expressions.

An even more integrated approach to classifying facial expressions requires multi-modal fusion through the amalgamation of sound, body posture and surrounding environmental data. These additional sensory sources build context, which is considered in the literature to be invaluable abstract information. We would expect that the use of sensory context would facilitate the ability to differentiate between expressions previously inseparable using a single medium such as a still image or video.

The conclusions above were obtained using a new classification framework introduced in this thesis and developed by the author. As mentioned in the introduction, the combination of multiple RFs as classifiers in the manner described in this thesis is a novel idea. The results presented here show that RF collections are the most significant aspect of the classifier and opens the door to further research on this topic.
The output of the RF collections\textsuperscript{22} was particularly significant as it served as a bridge between the preceding data analysis stage and the subsequent labelling stage. This output is intriguing since each of its elements originates from an accurate individual RF that is highly specialised to identify the presence or absence of just one state\textsuperscript{23}. Making sense of this output is a challenge in and of itself, approached here using an SVM labeller. Identifying the optimal tool for manipulation of the RF collection output is one of the more abstract areas of further work.

The appearance-based approach was used in this thesis. We believe that this choice enhanced our ability to achieve a consistent mapping from an input video to a labelled output expression. The general, but relatively slow, trend in the research community towards the use of videos for training the classifier, as well as the above observations, supports a paradigm shift towards video-based frameworks for facial expression recognition.

\textsuperscript{22} Named the poPDF vector here.

\textsuperscript{23} In this project, the ‘states’ are the different expressions.
Appendix I References


41. Fleet, D., *Model-Based Human Pose Tracking*, in *IEEE Workshop on Evaluation of Articulated Human Motion and Pose Estimation* 2007: Minneapolis, Minnesota, USA.


Appendix II  Methodology and Implementation Flowcharts

Flowchart 1: Training Database Preparation
Flowchart 2: Feature analysis and dimensionality reduction
Generate Training Labels according to 3 Schemes p. 56

Training Labels

Generate 3 Yaw Specific RF Collections for each Feature p. 55

Evaluate the FAR-CRIS using RF Collections to produce CRIS-poPDF

CRIS-poPDF

Train SVM and k-NN labellers on ‘training’ half of CRIS-poPDF p. 57

Flowchart 3: Classifier training
Input Image

Face Detection using PittPatt  

Crop face according to yaw bin using detected landmarks  

Resize cropped face according to yaw bin  

Smooth Image and Normalise Histogram  

Analyse face using feature operator  

Reshape feature image into feature vector  

Project feature vector on feature and yaw specific Eigenspace  

Evaluate face using feature and yaw specific RF collection to produce poPDF  

Evaluate poPDF using SVM or k-NN to obtain label  

Flowchart 4: Single test image classification
## Appendix III  Sample Feature Images

<table>
<thead>
<tr>
<th>Input</th>
<th>CRIS Image</th>
<th>Original BU-3DFE Image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1" alt="CRIS Image" /></td>
<td><img src="image2" alt="Original BU-3DFE Image" /></td>
</tr>
<tr>
<td>Sobel</td>
<td><img src="image3" alt="CRIS Image" /></td>
<td><img src="image4" alt="Original BU-3DFE Image" /></td>
</tr>
</tbody>
</table>
Table 22: Sample image from the CRIS (left) and from the original BU-3DFE database (right) with their respective S, DL and LBP feature images. The full sized texture images from the BU-3DFE database were included here simply for visual purposes.
Appendix IV  Mean Feature Images

Mean Sobel Feature Images

<table>
<thead>
<tr>
<th>Bin</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td>Disgust</td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
</tr>
<tr>
<td>Fear</td>
<td><img src="image25.png" alt="Image" /></td>
<td><img src="image26.png" alt="Image" /></td>
<td><img src="image27.png" alt="Image" /></td>
<td><img src="image28.png" alt="Image" /></td>
<td><img src="image29.png" alt="Image" /></td>
<td><img src="image30.png" alt="Image" /></td>
<td><img src="image31.png" alt="Image" /></td>
<td><img src="image32.png" alt="Image" /></td>
<td><img src="image33.png" alt="Image" /></td>
<td><img src="image34.png" alt="Image" /></td>
<td><img src="image35.png" alt="Image" /></td>
<td><img src="image36.png" alt="Image" /></td>
</tr>
<tr>
<td>Happiness</td>
<td><img src="image37.png" alt="Image" /></td>
<td><img src="image38.png" alt="Image" /></td>
<td><img src="image39.png" alt="Image" /></td>
<td><img src="image40.png" alt="Image" /></td>
<td><img src="image41.png" alt="Image" /></td>
<td><img src="image42.png" alt="Image" /></td>
<td><img src="image43.png" alt="Image" /></td>
<td><img src="image44.png" alt="Image" /></td>
<td><img src="image45.png" alt="Image" /></td>
<td><img src="image46.png" alt="Image" /></td>
<td><img src="image47.png" alt="Image" /></td>
<td><img src="image48.png" alt="Image" /></td>
</tr>
<tr>
<td>Neutrality</td>
<td><img src="image49.png" alt="Image" /></td>
<td><img src="image50.png" alt="Image" /></td>
<td><img src="image51.png" alt="Image" /></td>
<td><img src="image52.png" alt="Image" /></td>
<td><img src="image53.png" alt="Image" /></td>
<td><img src="image54.png" alt="Image" /></td>
<td><img src="image55.png" alt="Image" /></td>
<td><img src="image56.png" alt="Image" /></td>
<td><img src="image57.png" alt="Image" /></td>
<td><img src="image58.png" alt="Image" /></td>
<td><img src="image59.png" alt="Image" /></td>
<td><img src="image60.png" alt="Image" /></td>
</tr>
<tr>
<td>Sadness</td>
<td><img src="image61.png" alt="Image" /></td>
<td><img src="image62.png" alt="Image" /></td>
<td><img src="image63.png" alt="Image" /></td>
<td><img src="image64.png" alt="Image" /></td>
<td><img src="image65.png" alt="Image" /></td>
<td><img src="image66.png" alt="Image" /></td>
<td><img src="image67.png" alt="Image" /></td>
<td><img src="image68.png" alt="Image" /></td>
<td><img src="image69.png" alt="Image" /></td>
<td><img src="image70.png" alt="Image" /></td>
<td><img src="image71.png" alt="Image" /></td>
<td><img src="image72.png" alt="Image" /></td>
</tr>
<tr>
<td>Surprise</td>
<td><img src="image73.png" alt="Image" /></td>
<td><img src="image74.png" alt="Image" /></td>
<td><img src="image75.png" alt="Image" /></td>
<td><img src="image76.png" alt="Image" /></td>
<td><img src="image77.png" alt="Image" /></td>
<td><img src="image78.png" alt="Image" /></td>
<td><img src="image79.png" alt="Image" /></td>
<td><img src="image80.png" alt="Image" /></td>
<td><img src="image81.png" alt="Image" /></td>
<td><img src="image82.png" alt="Image" /></td>
<td><img src="image83.png" alt="Image" /></td>
<td><img src="image84.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Table 23: Mean Sobel feature images for the 7 expressions in the 12 yaw bins

---

24 Feature images have been resized and normalised for viewing and formatting purposes. The dimensions used for each yaw bin are shown in Table 7 on p. 47.
### Table 24: Mean Laplace feature images for the 7 expressions in the 12 yaw bins

<table>
<thead>
<tr>
<th>Bin</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td>Disgust</td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
<td><img src="image21" alt="Image" /></td>
<td><img src="image22" alt="Image" /></td>
<td><img src="image23" alt="Image" /></td>
<td><img src="image24" alt="Image" /></td>
</tr>
<tr>
<td>Fear</td>
<td><img src="image25" alt="Image" /></td>
<td><img src="image26" alt="Image" /></td>
<td><img src="image27" alt="Image" /></td>
<td><img src="image28" alt="Image" /></td>
<td><img src="image29" alt="Image" /></td>
<td><img src="image30" alt="Image" /></td>
<td><img src="image31" alt="Image" /></td>
<td><img src="image32" alt="Image" /></td>
<td><img src="image33" alt="Image" /></td>
<td><img src="image34" alt="Image" /></td>
<td><img src="image35" alt="Image" /></td>
<td><img src="image36" alt="Image" /></td>
</tr>
<tr>
<td>Happiness</td>
<td><img src="image37" alt="Image" /></td>
<td><img src="image38" alt="Image" /></td>
<td><img src="image39" alt="Image" /></td>
<td><img src="image40" alt="Image" /></td>
<td><img src="image41" alt="Image" /></td>
<td><img src="image42" alt="Image" /></td>
<td><img src="image43" alt="Image" /></td>
<td><img src="image44" alt="Image" /></td>
<td><img src="image45" alt="Image" /></td>
<td><img src="image46" alt="Image" /></td>
<td><img src="image47" alt="Image" /></td>
<td><img src="image48" alt="Image" /></td>
</tr>
<tr>
<td>Neutrality</td>
<td><img src="image49" alt="Image" /></td>
<td><img src="image50" alt="Image" /></td>
<td><img src="image51" alt="Image" /></td>
<td><img src="image52" alt="Image" /></td>
<td><img src="image53" alt="Image" /></td>
<td><img src="image54" alt="Image" /></td>
<td><img src="image55" alt="Image" /></td>
<td><img src="image56" alt="Image" /></td>
<td><img src="image57" alt="Image" /></td>
<td><img src="image58" alt="Image" /></td>
<td><img src="image59" alt="Image" /></td>
<td><img src="image60" alt="Image" /></td>
</tr>
<tr>
<td>Sadness</td>
<td><img src="image61" alt="Image" /></td>
<td><img src="image62" alt="Image" /></td>
<td><img src="image63" alt="Image" /></td>
<td><img src="image64" alt="Image" /></td>
<td><img src="image65" alt="Image" /></td>
<td><img src="image66" alt="Image" /></td>
<td><img src="image67" alt="Image" /></td>
<td><img src="image68" alt="Image" /></td>
<td><img src="image69" alt="Image" /></td>
<td><img src="image70" alt="Image" /></td>
<td><img src="image71" alt="Image" /></td>
<td><img src="image72" alt="Image" /></td>
</tr>
<tr>
<td>Surprise</td>
<td><img src="image73" alt="Image" /></td>
<td><img src="image74" alt="Image" /></td>
<td><img src="image75" alt="Image" /></td>
<td><img src="image76" alt="Image" /></td>
<td><img src="image77" alt="Image" /></td>
<td><img src="image78" alt="Image" /></td>
<td><img src="image79" alt="Image" /></td>
<td><img src="image80" alt="Image" /></td>
<td><img src="image81" alt="Image" /></td>
<td><img src="image82" alt="Image" /></td>
<td><img src="image83" alt="Image" /></td>
<td><img src="image84" alt="Image" /></td>
</tr>
</tbody>
</table>
Mean Local Binary Pattern Feature Images

<table>
<thead>
<tr>
<th>Bin</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td>Disgust</td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
</tr>
<tr>
<td>Fear</td>
<td><img src="image25.png" alt="Image" /></td>
<td><img src="image26.png" alt="Image" /></td>
<td><img src="image27.png" alt="Image" /></td>
<td><img src="image28.png" alt="Image" /></td>
<td><img src="image29.png" alt="Image" /></td>
<td><img src="image30.png" alt="Image" /></td>
<td><img src="image31.png" alt="Image" /></td>
<td><img src="image32.png" alt="Image" /></td>
<td><img src="image33.png" alt="Image" /></td>
<td><img src="image34.png" alt="Image" /></td>
<td><img src="image35.png" alt="Image" /></td>
<td><img src="image36.png" alt="Image" /></td>
</tr>
<tr>
<td>Happiness</td>
<td><img src="image37.png" alt="Image" /></td>
<td><img src="image38.png" alt="Image" /></td>
<td><img src="image39.png" alt="Image" /></td>
<td><img src="image40.png" alt="Image" /></td>
<td><img src="image41.png" alt="Image" /></td>
<td><img src="image42.png" alt="Image" /></td>
<td><img src="image43.png" alt="Image" /></td>
<td><img src="image44.png" alt="Image" /></td>
<td><img src="image45.png" alt="Image" /></td>
<td><img src="image46.png" alt="Image" /></td>
<td><img src="image47.png" alt="Image" /></td>
<td><img src="image48.png" alt="Image" /></td>
</tr>
<tr>
<td>Neutrality</td>
<td><img src="image49.png" alt="Image" /></td>
<td><img src="image50.png" alt="Image" /></td>
<td><img src="image51.png" alt="Image" /></td>
<td><img src="image52.png" alt="Image" /></td>
<td><img src="image53.png" alt="Image" /></td>
<td><img src="image54.png" alt="Image" /></td>
<td><img src="image55.png" alt="Image" /></td>
<td><img src="image56.png" alt="Image" /></td>
<td><img src="image57.png" alt="Image" /></td>
<td><img src="image58.png" alt="Image" /></td>
<td><img src="image59.png" alt="Image" /></td>
<td><img src="image60.png" alt="Image" /></td>
</tr>
<tr>
<td>Sadness</td>
<td><img src="image61.png" alt="Image" /></td>
<td><img src="image62.png" alt="Image" /></td>
<td><img src="image63.png" alt="Image" /></td>
<td><img src="image64.png" alt="Image" /></td>
<td><img src="image65.png" alt="Image" /></td>
<td><img src="image66.png" alt="Image" /></td>
<td><img src="image67.png" alt="Image" /></td>
<td><img src="image68.png" alt="Image" /></td>
<td><img src="image69.png" alt="Image" /></td>
<td><img src="image70.png" alt="Image" /></td>
<td><img src="image71.png" alt="Image" /></td>
<td><img src="image72.png" alt="Image" /></td>
</tr>
<tr>
<td>Surprise</td>
<td><img src="image73.png" alt="Image" /></td>
<td><img src="image74.png" alt="Image" /></td>
<td><img src="image75.png" alt="Image" /></td>
<td><img src="image76.png" alt="Image" /></td>
<td><img src="image77.png" alt="Image" /></td>
<td><img src="image78.png" alt="Image" /></td>
<td><img src="image79.png" alt="Image" /></td>
<td><img src="image80.png" alt="Image" /></td>
<td><img src="image81.png" alt="Image" /></td>
<td><img src="image82.png" alt="Image" /></td>
<td><img src="image83.png" alt="Image" /></td>
<td><img src="image84.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Table 25: Mean Local Binary Pattern feature images for the 7 expressions in the 12 yaw bins
Appendix V  Sobel kNN-Labelling Classifier Accuracies

Figure 33: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 100 trees per forest under the two different labelling schemes using Sobel features

Figure 34: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 100 trees per forest under the two different labelling schemes using Sobel features

Figure 35: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 100 trees per forest using Sobel features
Figure 36: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 200 trees per forest under the two different labelling schemes using Sobel features.

Figure 37: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 200 trees per forest under the two different labelling schemes using Sobel features.

Figure 38: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 200 trees per forest using Sobel features.
Figure 39: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 300 trees per forest under the two different labelling schemes using Sobel features.

Figure 40: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 300 trees per forest under the two different labelling schemes using Sobel features.

Figure 41: k-NN labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 300 trees per forest using Sobel features.
Appendix VI  Discrete Laplace SVM-Labelling Classifier

Accuracies

Figure 42: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 200 trees per forest under the two different labelling schemes using Discrete Laplace features

Figure 43: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 200 trees per forest under the two different labelling schemes using Discrete Laplace features

Figure 44: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 200 trees per forest using Discrete Laplace features
Figure 45: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 300 trees per forest under the two different labelling schemes using Discrete Laplace features.

Figure 46: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 300 trees per forest under the two different labelling schemes using Discrete Laplace features.

Figure 47: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 300 trees per forest using Discrete Laplace features.

<table>
<thead>
<tr>
<th>Forest Type</th>
<th>Number of Trees</th>
<th>Labelling Scheme 1</th>
<th>Labelling Scheme 2</th>
<th>Labelling Scheme 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>53.3136</td>
<td>75.4200</td>
<td>75.5134</td>
</tr>
<tr>
<td>1</td>
<td>300</td>
<td>54.3591</td>
<td>76.0983</td>
<td>76.1699</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>58.2327</td>
<td>75.4947</td>
<td>75.5009</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>59.1848</td>
<td>76.4343</td>
<td>76.0019</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>--</td>
<td>--</td>
<td>70.6596</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>--</td>
<td>--</td>
<td>70.9894</td>
</tr>
</tbody>
</table>

Table 26: Central bins (5, 6, 7 and 8) mean classification accuracies for Discrete Laplace feature Classifiers with SVM labelling (all values are %)
Appendix VII  Local Binary Pattern SVM-Labelling Classifier Accuracies

Figure 48: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 200 trees per forest under the two different labelling schemes using Local Binary Pattern features

Figure 49: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 200 trees per forest under the two different labelling schemes using Local Binary Pattern features

Figure 50: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 200 trees per forest using Local Binary Pattern features
Figure 51: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 1’ RF collection with 300 trees per forest under the two different labelling schemes using Local Binary Pattern features.

Figure 52: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 2’ RF collection with 300 trees per forest under the two different labelling schemes using Local Binary Pattern features.

Figure 53: SVM labelling accuracy of the test half of the CRIS-poPDF for a ‘type 3’ RF collection with 300 trees per forest using Local Binary Pattern features.

<table>
<thead>
<tr>
<th>Forest Type</th>
<th>Number of Trees</th>
<th>Labelling Scheme 1</th>
<th>Labelling Scheme 2</th>
<th>Labelling Scheme 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>52.4860</td>
<td>77.5264</td>
<td>77.6914</td>
</tr>
<tr>
<td>1</td>
<td>300</td>
<td>53.1301</td>
<td>77.9745</td>
<td>77.4673</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>56.0081</td>
<td>77.3086</td>
<td>77.9932</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>56.8046</td>
<td>77.8407</td>
<td><strong>78.9172</strong></td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>--</td>
<td>--</td>
<td>73.1052</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>--</td>
<td>--</td>
<td>73.7866</td>
</tr>
</tbody>
</table>

Table 27: Central bins (5, 6, 7 and 8) mean classification accuracies for Local Binary Pattern feature classifiers with SVM labelling (all values are %)
Appendix VIII  Still Image Classification Results

The classification accuracy plots provided in this appendix are arranged in the order listed in Table 2 on p. 34. For each database, the first plot shows the number of images detected in each yaw bin (see Figure 10 on p. 48 for the yaw distribution) for each expression. The remaining three plots for each database pertain to the accuracy of classification of the images for using Sobel, discrete Laplace and Local Binary Pattern classifiers respectively. The accuracy plots also contain the yaw bin information.

Karolinska Directed Emotional Faces (KDEF) Database

![Figure 54: Number of face images found per yaw bin according to true expression label for KDEF](image)

![Figure 55: Classification accuracy of KDEF using the optimal Sobel based classifier from Table 13](image)
Figure 56: Classification accuracy of KDEF using the optimal Discrete Laplace based classifier from Table 13

Figure 57: Classification accuracy of KDEF using the optimal LBP based classifier from Table 13
Radboud Faces Database (RaFD) [95]

Figure 58: Number of face images found per yaw bin according to true expression label for RaFD

Figure 59 Classification accuracy of RaFD using the optimal Sobel based classifier from Table 13
Figure 60: Classification accuracy of RaFD using the optimal Discrete Laplace based classifier from Table 13

Figure 61: Classification accuracy of RaFD using the optimal LBP based classifier from Table 13
Figure 62: Number of face images found per yaw bin according to true expression label for JAFFE

Figure 63: Classification accuracy of JAFFE using the optimal Sobel based classifier from Table 13
Figure 64: Classification accuracy of JAFFE using the optimal Discrete Laplace based classifier from Table 13

Figure 65: Classification accuracy of JAFFE using the optimal LBP based classifier from Table 13
Figure 66: Number of face images found per yaw bin according to true expression label for PICS-pain

Figure 67: Classification accuracy of PICS-pain using the optimal Sobel based classifier from Table 13
Figure 68: Classification accuracy of PICS-pain using the optimal Discrete Laplace based classifier from Table 13

Figure 69: Classification accuracy of PICS-pain using the optimal LBP based classifier from Table 13
Static Facial Expressions In The Wild (SFEW) [98]

Figure 70: Number of face images found per yaw bin according to true expression label for SFEW

Figure 71: Classification accuracy of SFEW using the optimal Sobel based classifier from Table 13

119
Figure 72: Classification accuracy of SFEW using the optimal Discrete Laplace based classifier from Table 13

Figure 73: Classification accuracy of SFEW using the optimal LBP based classifier from Table 13
Appendix IX  Video Classification Results

The classification accuracy plots provided in this appendix are arranged in the order listed in Table 2 on p. 34. For each database, the first plot shows the number of videos analysed for each expression. The remaining three plots for each database pertain to the accuracy of classification of the videos using Sobel, discrete Laplace and Local Binary Pattern classifiers respectively. The accuracy plots show results for labelling schemes 2 and 3 from p. 58.

Binghamton University 3D Dynamic Facial Expression Database (BU-4DFE) [90]

![Figure 74: Number of video sequences tested according to true expression label for BU-4DFE](image)

![Figure 75: Classification accuracy of BU-4DFE using the optimal Sobel based classifier (with labelling schemes 2 and 3) from Table 13](image)
Figure 76: Classification accuracy of BU-4DFE using the optimal Discrete Laplace based classifier (with labelling schemes 2 and 3) from Table 13

Figure 77: Classification accuracy of BU-4DFE using the optimal LBP based classifier (with labelling schemes 2 and 3) from Table 13
Facial Expressions and Emotion Database (FEED)

Figure 78: Number of video sequences tested according to true expression label for FEED

Figure 79: Classification accuracy of FEED using the optimal Sobel based classifier (with labelling schemes 2 and 3) from Table 13
Figure 80: Classification accuracy of FEED using the optimal Discrete Laplace based classifier (with labelling schemes 2 and 3) from Table 13

Figure 81: Classification accuracy of FEED using the optimal LBP based classifier (with labelling schemes 2 and 3) from Table 13
Acted Facial Expressions In The Wild (AFEW) Database [100]

Figure 82: Number of video sequences tested according to true expression label for AFEW

Figure 83: Classification accuracy of AFEW using the optimal Sobel based classifier (with labelling schemes 2 and 3) from Table 13
Figure 84: Classification accuracy of AFEW using the optimal Discrete Laplace based classifier (with labelling schemes 2 and 3) from Table 13

Figure 85: Classification accuracy of AFEW using the optimal LBP based classifier (with labelling schemes 2 and 3) from Table 13
Appendix X  Numerical Confusion Matrices for BU-3DFE

Table 28: Numerical values related to the Sobel confusion matrix in Figure 24 (all values are %)

<table>
<thead>
<tr>
<th>True Expression</th>
<th>AN</th>
<th>DI</th>
<th>FE</th>
<th>HA</th>
<th>NE</th>
<th>SA</th>
<th>SU</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AN</td>
<td>75.84</td>
<td>5.36</td>
<td>3.33</td>
<td>1.01</td>
<td>4.22</td>
<td>8.40</td>
<td>1.84</td>
<td>5430</td>
</tr>
<tr>
<td>DI</td>
<td>7.57</td>
<td>80.17</td>
<td>4.14</td>
<td>2.55</td>
<td>0.95</td>
<td>2.64</td>
<td>1.99</td>
<td>5380</td>
</tr>
<tr>
<td>FE</td>
<td>4.28</td>
<td>5.06</td>
<td>72.05</td>
<td>6.25</td>
<td>2.93</td>
<td>5.50</td>
<td>3.92</td>
<td>5181</td>
</tr>
<tr>
<td>HA</td>
<td>1.70</td>
<td>2.55</td>
<td>6.15</td>
<td>85.48</td>
<td>1.67</td>
<td>1.53</td>
<td>0.92</td>
<td>5220</td>
</tr>
<tr>
<td>NE</td>
<td>3.80</td>
<td>0.83</td>
<td>2.89</td>
<td>1.32</td>
<td>82.22</td>
<td>6.04</td>
<td>2.89</td>
<td>1209</td>
</tr>
<tr>
<td>SA</td>
<td>9.59</td>
<td>2.54</td>
<td>4.66</td>
<td>1.33</td>
<td>5.91</td>
<td>73.11</td>
<td>2.86</td>
<td>5109</td>
</tr>
<tr>
<td>SU</td>
<td>2.08</td>
<td>2.54</td>
<td>4.27</td>
<td>1.60</td>
<td>3.30</td>
<td>3.77</td>
<td>82.43</td>
<td>4611</td>
</tr>
</tbody>
</table>

Table 29: Numerical values related to the discrete Laplace confusion matrix in Figure 24 (all values are %)

<table>
<thead>
<tr>
<th>True Expression</th>
<th>AN</th>
<th>DI</th>
<th>FE</th>
<th>HA</th>
<th>NE</th>
<th>SA</th>
<th>SU</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AN</td>
<td>76.70</td>
<td>5.62</td>
<td>2.97</td>
<td>1.38</td>
<td>4.16</td>
<td>7.53</td>
<td>1.64</td>
<td>5430</td>
</tr>
<tr>
<td>DI</td>
<td>6.25</td>
<td>81.12</td>
<td>4.00</td>
<td>2.55</td>
<td>1.17</td>
<td>2.64</td>
<td>2.29</td>
<td>5380</td>
</tr>
<tr>
<td>FE</td>
<td>3.65</td>
<td>4.65</td>
<td>72.55</td>
<td>7.62</td>
<td>2.28</td>
<td>4.73</td>
<td>4.52</td>
<td>5181</td>
</tr>
<tr>
<td>HA</td>
<td>1.30</td>
<td>3.51</td>
<td>7.80</td>
<td>82.34</td>
<td>1.46</td>
<td>2.16</td>
<td>1.44</td>
<td>5220</td>
</tr>
<tr>
<td>NE</td>
<td>3.23</td>
<td>0.66</td>
<td>2.56</td>
<td>1.74</td>
<td>83.71</td>
<td>5.13</td>
<td>2.98</td>
<td>1209</td>
</tr>
<tr>
<td>SA</td>
<td>6.83</td>
<td>2.82</td>
<td>4.83</td>
<td>1.49</td>
<td>5.32</td>
<td>76.39</td>
<td>2.31</td>
<td>5109</td>
</tr>
<tr>
<td>SU</td>
<td>2.19</td>
<td>2.67</td>
<td>3.67</td>
<td>1.95</td>
<td>3.25</td>
<td>2.49</td>
<td>83.78</td>
<td>4611</td>
</tr>
</tbody>
</table>

Table 30: Numerical values related to the local binary pattern confusion matrix in Figure 24 (all values are %)

<table>
<thead>
<tr>
<th>True Expression</th>
<th>AN</th>
<th>DI</th>
<th>FE</th>
<th>HA</th>
<th>NE</th>
<th>SA</th>
<th>SU</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AN</td>
<td>76.70</td>
<td>5.62</td>
<td>2.97</td>
<td>1.38</td>
<td>4.16</td>
<td>7.53</td>
<td>1.64</td>
<td>5430</td>
</tr>
<tr>
<td>DI</td>
<td>6.25</td>
<td>81.12</td>
<td>4.00</td>
<td>2.55</td>
<td>1.17</td>
<td>2.64</td>
<td>2.29</td>
<td>5380</td>
</tr>
<tr>
<td>FE</td>
<td>3.65</td>
<td>4.65</td>
<td>72.55</td>
<td>7.62</td>
<td>2.28</td>
<td>4.73</td>
<td>4.52</td>
<td>5181</td>
</tr>
<tr>
<td>HA</td>
<td>1.30</td>
<td>3.51</td>
<td>7.80</td>
<td>82.34</td>
<td>1.46</td>
<td>2.16</td>
<td>1.44</td>
<td>5220</td>
</tr>
<tr>
<td>NE</td>
<td>3.23</td>
<td>0.66</td>
<td>2.56</td>
<td>1.74</td>
<td>83.71</td>
<td>5.13</td>
<td>2.98</td>
<td>1209</td>
</tr>
<tr>
<td>SA</td>
<td>6.83</td>
<td>2.82</td>
<td>4.83</td>
<td>1.49</td>
<td>5.32</td>
<td>76.39</td>
<td>2.31</td>
<td>5109</td>
</tr>
<tr>
<td>SU</td>
<td>2.19</td>
<td>2.67</td>
<td>3.67</td>
<td>1.95</td>
<td>3.25</td>
<td>2.49</td>
<td>83.78</td>
<td>4611</td>
</tr>
</tbody>
</table>
Appendix XI  Random Forests

Leo Breiman introduced Random Forests (RFs) in their current form in a 2001 article [10]. The bulk of his inspiration came from Amit and Geman’s decision tree handwriting recognition work [151] and Ho’s segmentation of the training set for random subspace based classifiers [152].

Earlier RF research such as Dietterich’s [153] investigation of random threshold selection at parent (decision) nodes and Breiman’s own work on altering training sample output values for regression purposes [154], resulted in favourable generalization errors results, particularly when classification speed was factored in. However, compared to state-of-the-art tools such as AdaBoost [155], these early RFs were on average, a step behind.

The reason for the shortcomings was the inappropriate selection of the randomised attributes. Effective vote-based classifiers require the correlation between the constituents (individual trees in this case) to remain low without influencing their collective strength. Breiman showed that randomly selected, or alternatively, randomly weighted linear combinations of certain features considered for splitting at each node, along with traditional training set bagging techniques, achieved the desired correlation condition.

Construction of each tree in a forest involves random selection of training vectors from the full training set, followed by random selection of features to consider for splitting at each decision node. During training, the features and corresponding threshold values selected to ultimately direct samples at each node maximise the information gain measure. This measure is used to evaluate how well a particular choice of feature-threshold pair would disperse training vectors of different classes that reached the node (also known as training vectors housed at a node). Furthermore, the measure used must accommodate the feature’s type (continuous or discrete) and range. Information gain measures rely on entropy minimisation following node splitting. However, in cases where features can take on a large range of values, the effectiveness of the pure information gain measure is diminished. Alternatively, an information gain ratio is at times employed to counteract the effects of large ranges of feature values. Another technique used to decide on feature-threshold pairs is random value selection. A threshold value is picked randomly from the range of training vectors housed at the node, removing any heuristic evaluation from the construction of the forest. However, Gashler et al. reported that random threshold selection did not perform quite as well as entropy based measures on large datasets containing ‘irrelevant attributes’ [156]. Gashler’s results could be extended to imply that an equilibrium, defined as a manageable, finite depth for each tree in the RF, is difficult to achieve. Thus, the use of a
heuristic seems particularly intuitive when considering that Breiman intended to keep the depth of the RFs uncapped.

As a classifier, RFs rely on large scale voting by groups of decision trees to estimate the posterior probability density function for a given observation vector. The core strength of the algorithm lies in the speed at which a high number of trees can be evaluated. A series of simple threshold comparisons at the decision nodes of each tree determines the local posterior distribution and is dependent on the training samples that reach the leaf node. Subsequent groupings of the posterior distributions across the entire forest(s) provide the global distribution.

Through experimentation, Breiman showed that RFs avoided systemic bias and overfitting, two pitfalls typical of any classifier. AdaBoost and other classifiers trained under an iterative sample weighting framework possess similar strengths. However, the weighting process itself can be seen as additive noise, thereby corrupting the training data. Although self-evaluation of each tree using the ‘out-of-bag’ samples provides a potential weighting scheme (misclassified samples could be weighted more heavily for the construction of the next tree), Breiman ignore this notion, and instead relies on the law of large numbers to eliminate such pitfalls. Furthermore, tree independence greatly facilitates parallelisation of both training (online and offline) and test stages in RFs. The former is strictly non-parallelisable when using AdaBoost classifiers based on decision trees. The significance of offline training speed increases with the growth of the training set. However, online (or live) training of a classifier is seldom possible and rarely parallelisable. Saffari et al. showed that online learning is not only feasible using RFs, but due to the tree independence criterion, also parallelisable [157].

The aforementioned inherent weighting scheme is one of several positive by-products of the bagging process applied to the training set. Unbiased estimates of the importance of different features, the strength of each tree, as well as correlation between different trees are obtained through the evaluation of the training samples not considered during the construction of each tree.

The most pertinent result to the construction of RFs is the importance of individual features in the classification process. By evaluating the out-of-bag samples for each tree and then altering specific feature values across the out-of-bag set and re-evaluating these same samples, a value for the importance of the altered feature can be obtained by averaging the classification accuracy ratio before and after alteration across the entire forest. The feature importance value could then be used to bias the selection of the features for splitting at each node, potentially reducing the size of each tree at the risk of doubling the full construction time.
Breiman’s original RF construction follows the formula outlined below:

- Given the complete training set $\mathcal{T}_r$ such that \( \{x_i \in \mathcal{T}_r: i = 1 \ldots N, |\mathcal{T}_r| = N \} \) and \( x_i = [f_1^i \ f_2^i \ldots \ f_m^i \ldots f_M^i]^\prime, \ |x_i| = M, \forall i \)

  Where \( f_m^i \) corresponds to the value of feature \( m \) in training vector \( x_i \).

- Given the complete label set $\mathcal{C}_l$ such that \( \{y_i \in \mathcal{C}_l: i = 1 \ldots N, |\mathcal{C}_l| = N \} \) where \( y_i \) corresponds to the label of training vector \( x_i \).

- Choose a value, \( m' \), where \( m' \ll M \), for the number of features randomly picked from \( 1 \ldots M \) without replacement and considered for the best split at each decision node of each tree. The resultant set of random features selected at node \( v \) of tree \( k \) is labelled \( m_v^k \).

Note: \( m' \) is kept constant throughout forest growth.

- Sample \( N \) training vectors from $\mathcal{T}_r$ with replacement, to form $\mathcal{T}_r^k$ (and corresponding $\mathcal{C}_l^k$): the training set used to construct the \( k \)-th tree,

\[
\begin{align*}
\{x_i^k \in \mathcal{T}_r^k: i' \in \{1 \ldots N\}, |\mathcal{T}_r^k| = N \} \\
\{y_i^k \in \mathcal{C}_l^k: i' \in \{1 \ldots N\}, |\mathcal{C}_l^k| = N \}
\end{align*}
\]

- With $\mathcal{T}_r^{v,k}$ representing the ‘remaining’ training set at node \( v \) of tree \( k \), we can say $\mathcal{T}_r_{1,k} = \mathcal{T}_r^k$.

Additions to the traditional algorithm we customised from the OpenCV [139] implementation of RFs used common decision tree construction norms:

- Requiring a minimum number of training samples \( |\mathcal{T}_r|_{\text{min}} \) to be housed at node \( v \) before splitting, such that

\[
\mathcal{T}_r^{v,k} \geq |\mathcal{T}_r|_{\text{min}} \forall v, k
\]

  This prevents unnecessarily complex trees where node splitting on a very small number of samples, a typical cause of overfitting in classifiers, is carried out.

- Requiring a minimum variance in the classes of samples housed at a node, or alternatively that no single class dominates $\mathcal{C}_l_v^k$. This is an approximation to usual classification tree algorithm’s condition whereby a leaf node is generated only when all the classes in the node are alike.

  By setting \( \%_{\mathcal{C}_l}^{\text{max}} \), the maximum percentage of samples with the same class at a node to a value above 95\%, a good approximation to node purity is achieved without the need to wait for the RF constructor to weed out the remaining 5\% of classes.

- Setting a maximal tree depth, \( D_{\text{max}} \), such that \( D_v \), the depth of node \( v \), follows

\[
D_v \leq D_{\text{max}} \forall v
\]

  By setting \( D_v \) to a significantly large value, for example, twice the average tree depth following uncapped construction, excessively deep paths that contribute little to the
classification process are eliminated. The assumption here is that other leaf node generating conditions would prevent tree growth before this constraint comes into play.

The conditions above do not result in a loss of information since the class distribution at each leaf node is still captured by $P_v(y|x)$.

Initialise *Forest* of size $K$ trees

```
for $k = 1 \ldots K$ {
    $T = \begin{bmatrix}
    (m^*, \theta^*)_v & v_L & v_R & P_v(y|x) \\
    \vdots & \vdots & \vdots & \vdots \\
    \end{bmatrix}$ \hspace{1em} // Tree structure
    $v_{\text{NEXT}} = 1$
    $D_v = 1$
    SplitNode($T^k$, $C^k$, $v_{\text{NEXT}}$, $D_v$)
    $Forest[k] = T$ // add the $k$th tree to the forest
}
```

SplitNode($T^k_v$, $C^k_v$, $v$, $D_v$) {
    
    $v_L = \infty$ \hspace{1em} // Initialise Left Child Node Address
    $v_R = \infty$ \hspace{1em} // Initialise Right Child Node Address
    $(m^*, \theta^*)_v = \infty$ \hspace{1em} // Initialise split feature and corresponding threshold

    if $|T^k_v| < |T^k_{\text{min}}| \land D_v > D_{\text{max}} \land C^k_v > C^k_{\text{max}}$
        Build normalized $P_v(y|x)$ using $C^k_v$
        Update $T$
        return
    end

    if $y_{i'} = y_{j'} \land y_{i'}, y_{j'} \in C^k_v$
        $P_v(y_{i'}|x) = 1$
        $P_v(y \neq y_{i'}|x) = 0$
        Update $T$
        return
    end

    else if $x_{i'} = x_{j'} \land x_{i'}, x_{j'} \in T^k_v$
        Build normalized $P_v(y|x)$ using $C^k_v$
        Update $T$
        return
    end

    Obtain $m^k_v$

    Evaluate the information gain criterion on $m^k_v$ to obtain maximising combination $(m^*, \theta^*)_v$

    $\forall x_{i'} \in T^k_v, i' = 1 \ldots |T^k_v|$
    
    if $f_{m^*}^{i'} \leq \theta^*$
        $x_{i'} \in T^k_{v_L}$
        $y_{i'} \in C^k_{v_{L}}$
    else
\[ x_i' \in \mathbb{T}_v^k \]
\[ y_i' \in \mathbb{C}_v^k \]
end

\[ v_L = LastNodeNumber + 1 \]
\[ v_R = v_L + 1 \]

SplitNode(\( \mathbb{T}_v^k \), \( \mathbb{C}_v^k \), \( v_L \), \( D_v + 1 \))
SplitNode(\( \mathbb{T}_v^k \), \( \mathbb{C}_v^k \), \( v_R \), \( D_v + 1 \))
\[ v_{\text{NEXT}} = v_R \]
Update \( T \)
return