FACIAL BEHAVIOUR ANALYSIS FOR CLINICAL APPLICATIONS

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

2019

By

Riyadh Almutiry
School of Computer Science
Contents

Abstract ................................................................. 12
Declaration .............................................................. 14
Copyright ................................................................. 15
Acknowledgements ....................................................... 17

1 Introduction .......................................................... 20
  1.1 Aims and Objectives ............................................. 21
  1.2 Contributions ................................................... 22
  1.3 Thesis structure .................................................. 22

2 Facial Feature Detection and Tracking Background ............... 24
  2.1 Statistical Shape Models (SSMs) .............................. 26
  2.2 Point Distribution Model PDM ................................. 28
  2.3 Modelling Shape Instances .................................... 29
  2.4 Fitting the Model to a New Image ............................. 31
  2.5 Statistical Appearance Models SAMs ......................... 32
  2.6 Active Shape Models (ASMs) .................................. 36
    2.6.1 Local Models for Each Point ............................. 37
  2.7 Active Appearance Models AAMs ............................. 39
  2.8 The Constrained Local Model ................................. 41
4.3.1 Emotion Theories ........................................ 89
4.3.2 Categorising Emotions ................................. 90
4.3.3 Multidimensional Space ............................... 92
4.4 Automatic Models of Facial Expression (AMFE) .... 93
  4.4.1 Challenges ........................................... 93
    4.4.1.1 Tracking Challenges ............................ 93
    4.4.1.2 Targeted Expressions ......................... 94
  4.4.2 Steps of Automatic Modelling of Facial Expression (AMFE) . 95
  4.4.3 Feature Extraction ................................. 97
    4.4.3.1 Gabor ....................................... 97
    4.4.3.2 Local Binary Patterns (LBP) .................. 98
4.5 Dataset Structure ....................................... 100
  4.5.1 Considerations ..................................... 100
  4.5.2 Collection Method ................................. 102
  4.5.3 Parameter Definitions .............................. 103
4.6 Analysis using Geometrical Features .................. 103
  4.6.1 Modelling Intra-subject variation .................. 103
  4.6.2 Modelling the Expression Subspace ............... 105
  4.6.3 Neutral Expression Distribution .................... 106
  4.6.4 Reducing Effect of Bias ............................ 108
  4.6.5 Onset and Offset Detection ......................... 109
  4.6.6 Refining Data .................................... 111
  4.6.7 Normalising Samples ............................... 112
4.7 Conclusion and Discussion ............................. 113

5 Gaze Estimation ........................................... 116
  5.1 Introduction ......................................... 116
  5.2 Eye tracking Methods .................................. 117
    5.2.1 Feature extraction .............................. 118
5.3 Challenges ..................................................... 120
5.4 Previous Work ............................................. 121
5.5 Summary .................................................... 126
5.6 Experiment ................................................ 128
  5.6.1 Feature Extraction .................................... 128
  5.6.2 Regression Forest ................................... 128
5.7 Conclusion and Discussion ............................. 130

6 Parkinson’s Disease Case Study .......................... 131
  6.1 Introduction ............................................. 131
  6.2 Clinical Assessment in Parkinson’s Disease ........ 133
  6.3 Facial Behaviour Research in Parkinson’s Disease .... 136
    6.3.1 Methods ........................................... 137
    6.3.2 Facial Expressivity Assessment in PD .......... 139
    6.3.3 Past studies ...................................... 140
    6.3.4 Summary .......................................... 147
  6.4 The Case Study .......................................... 150
    6.4.1 Data collection .................................... 151
    6.4.2 Annotation ........................................ 153
    6.4.3 Tracking Facial Features ......................... 153
    6.4.4 Irrelevant Expression Data Removal ............ 155
    6.4.5 Geometrical shape analysis ...................... 158
    6.4.6 Observation and Evaluation ..................... 165
    6.4.7 Textural-based Features ......................... 170
  6.5 Correlations ............................................ 171
  6.6 Discussion and Conclusion ........................... 182

7 Discussions and Future Work ............................ 185
  7.1 Challenges and Future work .......................... 187
List of Tables

3.1 Samples of several public datasets used in this project . . . . . . . . . . 52
3.2 Several public datasets for facial images . . . . . . . . . . . . . . . . . 53
3.3 Examined values of various parameters for 6 and 68 points RFCLM . . 55
3.4 Optimal RFCLM for First Level in stage 1 (6 points) . . . . . . . . . . 59
3.5 Effects of using the first best RFCLM as a second level . . . . . . . . 61
3.6 Effects of using a second level . . . . . . . . . . . . . . . . . . . . . . . 63
3.7 Best Performance of RFCLM for First Level in stage 2 (68 points) . . 66
3.8 Distances less than 10% of reference length. I-E refers to Initialised
   Error Distance either in a relative distance or pixels (px) . . . . . . . . 72
3.9 Top RFCLM under video tracking settings. . . . . . . . . . . . . . . . 72
3.10 A comparison with several state-of-the-art methods on 300W dataset,
    which is divided into two subsets: common subset (containing both HE-
    LEN and LFPW teset subsets) and challenging subset (IBUG dataset) . . 75
3.11 Examined AAM configuration parameters . . . . . . . . . . . . . . . . 77
3.12 Top Models for 4-points AAMs . . . . . . . . . . . . . . . . . . . . . . 82
3.13 Top Models 27-points AAMs . . . . . . . . . . . . . . . . . . . . . . . . 82
4.1 A set of pair Gabor kernels with sinusoidal functions. Each pair of
    images below represent Gabor function with a sin and cosine function
    respectively. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
4.2 Mardia normality test . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108
5.1 Statistical evaluation of gaze predictors with different input features ... 129

6.1 Description of PD case studies’ dataset. The order of item in *Facial expression nature* row correspond respectively to items in *Emotional classification* and *No. of diff. Expressions* rows ... 148

6.2 Summary of PD case studies analysis methods ... 149

6.3 Summary of subjects demographics and screening tests ... 150

6.4 Dataset summary showing total numbers of displayed expressions for each group ... 151

6.5 Collected controls participants dataset. ... 160

6.6 Overview of controls dataset with respect to each expression ... 161

6.7 Collected PD participants dataset. ... 162

6.8 Overview of PD dataset with respect to each expression ... 163

6.9 Effect of varying several modes within ±3 standard deviations for several facial expressions. ... 164

6.10 Least overlapping distributions of examined measures between *Healthy Controls* (HC) and *Parkinson's Disease* (PD). The table shows measures derived from both 27- and 51-point based features. ... 168

6.11 Correlations between QoL scores and subject-dependant models ... 172

6.12 Correlations between QoL scores and expression-based models ... 173

6.13 Normality test of Residuals ... 182

6.14 Normality test of Residuals ... 182
### List of Figures

2.1 Face image with 68 landmarks .................................................................. 28
2.2 PCA axes when applied to a set of 2D vectors [1] ................................. 29
2.3 Mean shape plot (red dots) with shape instance (black dots) .............. 30
2.4 Different modes of variation ................................................................. 31
2.5 Profile search along normal to model boundary; the red dot shows the current profile position in the search. ............................................. 38

3.1 Distance error is normalised by the outer-ocular distance only when optimising RFCLM. Final evaluations on 300W datasets were normalised by the inter-ocular distance for comparable results. .......................... 54
3.2 Selected RFCLM parameters to optimise ............................................ 54
3.3 First and Second stages of RFCLM ..................................................... 55
3.4 Top 6-points RFCLM models evaluated under various initialised errors 58
3.5 Top two sequence of RFCLMs (second model fixed) ......................... 60
3.6 Top 6-points RFCLM models evaluated under various initialised errors 62
3.7 Top 68-points RFCLM models evaluated under various initialised errors 65
3.8 Examples of initialised points using bounding boxes ........................... 67
3.9 RFCLM cost image ............................................................................ 69
3.10 A sample image from DISFA[2] dataset showing annotation of 66 Landmarks ..................................................................................... 70
3.11 Top RFCLM Tracking Results on DISFA dataset when used with a tracking threshold set to the 95% percentile value of the quality-of-fit. 71
3.12 51 points RFCLM comparison with results of 300W challenge [3] on indoor and outdoor images of 300W-2 dataset. 73
3.13 51 points RFCLM comparison with results of 300W challenge [3] on indoor images of 300W-2 dataset. 74
3.14 51 points RFCLM comparison with results of 300W challenge [3] on outdoor images of 300W-2 dataset. 75
3.15 An example of 27 facial landmarking scheme used for annotating samples of AAM experiment. First stage AAM are used to locate 4 points (in red). 78
3.16 Top subject-specific AAM models based on median at first stage. 79
3.17 Top subject-specific AAM models based on median at second stage. 80
3.18 Top subject-specific AAM models based on median at third stage. 81
4.1 Plutchik’s (1980) wheels of emotions[4]. 92
4.2 Process pipeline for expression modelling. 96
4.3 Local Binary Pattern (LBP) Feature Extraction using 3x3 Neighbourhood 99
4.4 LBP with the circular neighbourhood of different sizes, reprinted from[5] 100
4.5 Dataset structure. 101
4.6 Neutral data projected in several expression subspaces. 107
4.7 Sample of expression model parameter variation over time. 110
4.8 Detection of expression onset and offset using various frequencies kernels. 111
4.9 Frames-normalised expression data. 113
4.10 A summary of facial behaviour analysis process. 115
5.1 Eye structure and visual and optical axes. 119
5.2 Taken from Wollaston (1824). 120
5.3 The cumulative distribution of gaze angular error of three different features. 129
6.1 Summary of data collection process. 152
6.2 Mean and standard deviations of tracked frames rates per session. First
and second rows represent, the 27-point and 51-point, tracking results
respectively, while the first and second columns show categorised results
per expression and subject, respectively . . . . . . . . . . . . . . . . . . 154
6.3 Quantified 27-point tracking failures . . . . . . . . . . . . . . . . . . 156
6.4 Quantified 51-point tracking failures . . . . . . . . . . . . . . . . . . 157
6.5 Facial expression variation of PD subjects (red) compared to controls
subjects (green) using first mode of Happy model. Units of y-axis cor-
respond to units of standard deviations. . . . . . . . . . . . . . . . . . 167
6.6 Expression distribution histograms . . . . . . . . . . . . . . . . . . . 169
6.7 An example of 27 sampled facial features . . . . . . . . . . . . . . . . 170
6.8 Diagnostic plots of regression model . . . . . . . . . . . . . . . . . . . 175
6.9 Diagnostic plots of regression model . . . . . . . . . . . . . . . . . . . 176
6.10 Diagnostic plots of regression model . . . . . . . . . . . . . . . . . . . 178
6.11 Diagnostic plots of regression model . . . . . . . . . . . . . . . . . . . 179
6.12 Diagnostic plots of anger mimicked regression model . . . . . . . . . 180
6.13 Diagnostic plots of anger mimicked regression model . . . . . . . . . 181
6.14 Significant subject-dependent models . . . . . . . . . . . . . . . . . . 182
6.15 Significant general models . . . . . . . . . . . . . . . . . . . . . . . . 182
Abstract

Facial Behaviour Analysis For Clinical Applications
Riyadh Almutiry
A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2019

This thesis explores whether facial feature tracking and facial behaviour analysis can be used to monitor the severity of disease. Some chronic and progressive diseases, such as Parkinson’s Disease and Schizophrenia, are associated with impaired facial expressions. Since there is no cure for the disease, current treatments aim to reduce the severity of the symptoms in order to improve patients’ quality of life.

Clinicians use measures of a patient’s facial behaviour (such as expressivity) when assessing their current state, but this is usually done infrequently, and is not practical for daily monitoring. In diseases such as Parkinson’s the severity of symptoms vary from one day to the next, so an evaluation on any particular day may not be representative of their longer term status.

These limitations have attracted many researchers to investigate the feasibility of developing automatic methods for measuring facial expressivity which could be used daily. However, reported findings are inconsistent and sometimes contradictory.

In this thesis, we investigate whether measures can be derived from the results of facial feature tracking which correlate with disease severity, with the aim of providing more information to help clinical assessments.

We examine data from a study of patients with Parkinson’s disease in which each participant was recorded whilst making a variety of different expressions in response to prompts from a computer program. By tracking the face through each video sequence we were able to monitor their facial movements. We explored a variety of
different parameters measuring behaviour (such as the intensity or duration of each expression) in order to identify whether (a) there was a measurable difference between behaviour of people with Parkinson’s compared to controls and (b) whether such parameters were correlated with a “Quality-of-Life” (QoL) score estimated from daily questionnaires.

We analysed facial movements with respect to a 27- and 51-point model. We found that there are differences in facial behaviour between controls and those with the disease; and that for some individuals there are some parameters which correlate with the QoL scores. Based on our experiments, we found several impaired and correlated facial expressions with QoL scores in both 27- and 51-point models. First, in correlation assessment with QoL scores, we found that mimicked expressions of anger were the most consistent across all models, while in facial expression impairments, happy and mimicked-happy were the most consistent in both models. The former, suggests that mimicked expressions of anger are more suited for the assessment of PD QoL, while happy and mimicked-happy expressions are more suited for highlighting facial expression impairments in PD.
Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University
Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses
Acknowledgements

I would like to express my gratitude to my supervisor Prof. Tim Cootes for his continuous support and guidance on this project. Without his amazing experience, and the fieldwork it would not have been possible to complete this PhD.

I would like to express my sincere thanks to Prof. Ellen Poliakoff, Prof. Alison Yung and Dr. Samuel Couth and their group for providing me with the data I used in this project.

I would like to express my thanks to my family especially my mother and my wife for their continuous support and inspirations to finish this project.
Acronyms

AAM  Active Appearance Model
AD  Alzheimer Disease
AMFE  Automatic Modelling of Facial Expression
ANOVA  Analysis of Variance
ASD  Autism Spectrum Disorders
CNN  Convolutional Neural Network
CPM  Component Process Model
EGT  Eye Gaze Tracking
FACS  Facial Action Coding System
FEEST  Expressions of Emotion: Stimuli and Tests
FER  Facial Expression Recognition
FFDT  Facial Feature Detection and Tracking
GDS  Geriatric Depression Scale
GLR  Greedy Linear Regression
HBNN  Hierarchical Bayesian Neural Networks
HC  Healthy Controls
HCI  Human-Computer Interaction
ICRP-IEB  Interpersonal Communication Rating Protocol for Individual Expressive Behavior
LASSO  Least Absolute Shrinkage and Selection Operator
LBP  Local Binary Pattern
MCI  Mild Cognitive Impairment
MDS  Movement Disorder Society
MDS-UPDRS  MDS revision of the Unified Parkinson’s Disease Rating Scale
MDS  Minimum Data Set
MFCC  Mel-Frequency Cepstral Coefficient
MMSE  Mini–Mental State Examination
PC  Principal Component
PD  Parkinson’s Disease
PDD  Parkinson’s Disease Dementia
PoG  Point of Gaze
QoL  Quality-of-Life
RFCLM  Random Forest Regression Voting Constrained Local Model
RFR  Random Forest Regression
SSM  Statistical Shape Model
SVM  Support Vector Machine
TNO  The Netherlands Optical Society
ToM  Theory of Mind
Chapter 1

Introduction

The human face represents a major channel for non-verbal signals such as facial expressions and gaze direction. These signals can be used to regulate interactions, communicate emotions, display comprehension, and express agreement or disagreement. There is a growing demand for automatic analysis and understanding of facial behaviour in general and more specifically for gaze and facial expression, to facilitate various applications in multiple domains such as computer vision [6], psychology Jack et al. [7], Nakashima et al. [8], neuroscience [9–12].

In computer vision, many researchers attempt to establish an automatic method for encoding/decoding of facial expression and gaze, in a regression or classification framework. For example, to enable understanding of patterns associated with different behaviour [6, 13]. In psychology, facial expressions and gaze have been used to improve understandings of cognitive processes. For example, Nakashima et al. [8] examined the effect of gaze direction on memory of faces with various emotional expressions. Their findings suggested that, memory of faces were affected by perceived intentions influenced by gaze.

Another research domain focuses on the assessment of facial behaviour abnormality associated with neurological disorders. Reports of several past studies on multiple chronic
1.1. AIMS AND OBJECTIVES

Brain diseases such as Parkinson’s Disease (PD)\cite{9, 10} and Schizophrenia \cite{11, 12} have shown that subjects with the disease have reduced facial expressivity compared to people without the disease. Such impairment of facial expression along with other symptoms associated with the disease can lead to social and psychological problems such as isolation and depression, impacting patient’s overall Quality-of-Life (QoL) \cite{14}.

Since there is no cure for the diseases, clinicians aim to reduce the severity of the symptoms in order to improve patient overall QoL. For example, in PD, Levodopa is a common drug used to reduce the symptoms. However, since PD is progressive, the severity of symptoms will increase over time and clinicians would have to increase the dosage intake. Inaccurate assessment of the disease severity can lead to improper dosage intake, which can lead to negative impact on a patient’s health. This involves high risk of increasing PD and many other physiological and psychological problems \cite{15}. This raises the need for an accurate assessment of the severity of PD symptoms.

There have been previous attempts to develop automatic methods for analysing facial behaviour to monitor disease progression but none of the proposed approaches are sufficient for daily monitoring in naturalistic settings. To the best of our knowledge this is the first attempt to develop an automatic method to objectify PD severity from facial dynamics using video sequences captured under naturalistic settings.

1.1 Aims and Objectives

1. Choosing and optimizing a robust facial feature tracker which can be used as the basis for other behavioural models;
2. Evaluating feasibility of estimating gaze directions from 2D images using a non-intrusive method with a single camera.
3. Optimising the parameters of a facial feature tracker to enable accurate monitoring of facial movements.
4. Collect and annotate databases of videos to study the facial movements of people with diseases
5. Analyse such videos in order to identify whether there are significant differences in facial behaviour in people with diseases
6. Analyse the videos to identify behaviour measures which can be used to estimate disease severity.

1.2 Contributions

- We report our finding of optimal parameters of RFCLM for both detecting and tracking facial features based on evaluation on several public datasets.
- We evaluate several features to model gaze direction. Our findings suggest that using only geometrical features such as shape parameters and eye-corner distances can improve gaze estimation. We evaluate the use of different features around the eyes in a regression framework for estimating gaze direction, showing that the best results come from a combination of geometric features.
- We present a fully automatic model which can describe and quantify facial expressivity using a variety of facial features extracted from video sequences in naturalistic settings. Our model can highlight potential discriminative parameters found between PD and controls in several facial expressions.
- Part of this work, to examine potential differences of PD and controls, has been published in [16]
- We also identify measures of facial behaviour that are correlated with patients’ QoL scores.

1.3 Thesis structure

In the next chapter, we discuss several techniques used related to facial feature detection and tracking with a main focus toward describing the development of RFCLM and
AAM models.

Chapter 3 describes our optimisation procedure for several parameters of RFCLM and AAM using extensive experiments.

Chapter 4 describes how we model facial behaviour dynamics and how we derive informative features.

Chapter 5 gives an overview of gaze direction estimation highlighting several state-of-the-art approaches. At the end of this chapter we show preliminary experimental results for evaluating several facial features to predict gaze direction.

In Chapter 6, we evaluate our model on a group of PD subjects and controls to see if our model can identify significant differences between PD and control subjects. We also examine the correlations of our model parameters with PD QoL scores and showed that our model did find significant correlated parameters for several subjects.

Finally, in Chapter 7, we summarise our findings and highlight challenges found in this project and outline extensions for future work.
Chapter 2

Facial Feature Detection and Tracking Background

The detection/localisation of facial landmarks, features, or points is also called face alignment, is the process of identifying the coordinates of a set of points in the face image, usually around key features such as mouth, eyes and nose. The detection of these points is a primary process for many computer vision applications such as Facial Expression Recognition (FER), Point of Gaze (PoG) estimation and emotion intensity estimation. It is also a preliminary step for many automatic facial behaviour analyses as mentioned in the previous chapter. Further discussion of facial expression and gaze applications are also addressed in Section 4.2 and Section 5.1, respectively).

Facial Feature Detection and Tracking (FFDT) is challenging due to the large number of variations that can be exhibited in facial images. Despite the vast number of studies that have been undertaken, many facial-feature detectors are still vulnerable to real-world conditions. These include subject-related variations (e.g., hair, skin, structure, age, gender and head-wearable items, such as glasses and hats) and image-related variations (e.g., image quality and illumination). Self-occlusion can also occur during social interactions, such as head rotation, raising the hands over the face and so on.
All of these variations and occlusions make FFDT task very challenging.

FFDT methods can be categorized into two categories: generative and discriminative [17]. Generative methods usually aim to establish a parametric model for both shape and appearance features and solve FFDT task as an optimisation problem, to minimise the difference between the model image and the target image in an iterative framework. However, in discriminative models, a direct mapping function between image features and the target point/s location is learned instead. These can be further classified into: cascaded regression, exemplar-based, graphical models and deep learning methods [18].

A popular family of generative models is the *Active Appearance Model* (AAM), originally developed by Cootes et al. [19]. There are two representations for AAMs, holistic (full face) and part-based. Part-based models are less sensitive to occlusions than holistic models and can produce very high and accurate results [17]. A review on how different basic appearance models works is given in Section 2.5, 2.6 and 2.7. A major issue of most AAM methods, is that they suffer from either poor generalisation to unseen images or inefficient fitting processes, especially in uncontrolled settings [17].

Recent reviews on FFDT methods show that significant developments have been made under uncontrolled settings [17, 18]. Most state-of-the-art methods have achieved less than 5% in average error rate for most testing datasets and less than 10% on recent challenging datasets.

Majority of state-of-the-art methods are shifting toward CNN-based techniques. For example, Feng et al. [20] proposed a regression-based CNN with a novel loss function named (Wing loss), which employs improving the balance of training samples with small to medium error range by resampling with rotation, scaling and translation. Xiao et al. [21] also proposed a recurrent attentive-refinement (RAR) network, where locations of facial landmarks are initialised with a softmax regression layer then refined sequentially at each recurrent stage to incorporate prior detections information.
in current search.

Another popular approach that has been successfully applied to many computer-vision applications was proposed in [22], which combines Random Forest Regression-based voting and Constrained Local Model (RFCLM). An important success factor of this approach is the learning efficiency and inferences [23].

Since RFCLM models are robust in terms of locating feature points, they can be used to feed other behaviour-tracking models, such as gaze estimators, facial feature expressions and even more sophisticated models.

To fully understand how RFCLM works, we review several related techniques such as statistical shape model and machine learning approaches in particular boosted regression used for local patch experts. We also provide a brief review of techniques such as AAM that consider texture and shape information, which can provide more informative input to the behaviour analysis.

### 2.1 Statistical Shape Models (SSMs)

Since facial shape is not static/rigid, deformable/flexible models are used to represent the shape deformations. A common and well established approach is the Statistical Shape Model (SSM) [24, 25], which is effective for limiting deformations to legitimate shapes that are similar to those found in the training set.

Models of shape are usually represented by a set of points called “landmarks”, which are drawn on the object’s surface to describe its outer structure. However, it is also possible to model the internal structure. In order to understand how the statistical shape model works, it is essential to define the term shape.

**Shape:** “all the geometrical information that remains when location, scale and rotational effects are filtered out from an object”. [26]
2.1. **STATISTICAL SHAPE MODELS (SSMS)**

This definition illustrates that if variation of two or more shapes occurs only on properties, such as “location, scale and rotation”, it means those shapes will be identical. However, if the variation occurs in other properties, it will affect the degree of similarity in the shapes.

Every shape in the training set is defined by a vector before the statistical shape analysis begins. In terms of a $d$ dimensional shape, the $x$ shape, which contains $n$ points, is defined by an $nd$ vector, such as $x = \{x_{11}, ..., x_{1d}, ..., x_{nd}\}$.

A key concept of statistical shape models is to produce a model that is able to capture the variations of shapes in a specific object class without losing model specificity [24]. The model uses a *mean shape* derived from the training set as the base shape. In order to generate the mean shape, certain pre-processing steps, such as a generalised Procrustes analysis (GPA) [27], must be applied. This includes aligning all shapes into a common coordinate system after filtering out the global transformation effects ‘scale, rotation and orientation’ to compute the mean shape. The goal of the alignment is to minimise the sum of square distances ( $D = \sum |x_i - \bar{x}|^2$ ) between the current shape and the mean shape. The GPA process is iterative, and in each iteration, all shapes are aligned with the mean shape in order to compute the new mean shape. This process is repeated until no significant changes are found in the new mean shape, at which point, convergence is declared.

Considering a set of shapes similar to those in Figure 2.1, which contains a face image marked with 68 ‘landmark’ points, the process of creating the mean shape can be summarised as follows:
After normalising all shapes in the initial iteration of the process so that $|x_i| = 1$, select one shape to be the mean shape and define a reference frame. Then,

1. Align all shapes to the mean shape.
2. Calculate the new mean of the aligned shapes.
3. Normalise the mean shape so $|\bar{x}| = 1$.
4. If no significant changes are observed between the old and new mean shape, convergence is declared and the mean shape is found, otherwise go back to 2.

### 2.2 Point Distribution Model PDM

Different alignment approaches can result in different distribution models for the same sets of points. When aligning two different shapes, $x$ and $x_i$, the shape will lie on a hypersphere because this approach imposes a scaling constraint: $|x| = 1$. This makes the variations in the points more vulnerable to non-linear change. As suggested by Cootes [1], to maintain the non-linearity changes in the point distribution to a minimum, all shapes are transformed into a tangent space, which requires taking the shape projection.
2.3. MODELLING SHAPE INSTANCES

$x_i$ from the original shape $x_p$ on a hyperplane in which $(x_p - x_i) \cdot x_p = 0$ or $x_i \cdot x_p = 1$. In other words, when points on the hyperplane are at some distance from the original points, the vector $x_p - x_i$ will be normal to $x_p$; otherwise, they will both lie at the intersection of the hyperplane, in which case, $x_i \cdot x_p = 1$.

2.3 Modelling Shape Instances

![PCA axes](image)

Figure 2.2: PCA axes when applied to a set of 2D vectors [1]

Variations on parts of the mean shape that represent any of the training examples can be approximated with fewer modes by applying PCA. Figure 2.2 shows an example of the approximation of $x$ to $x'$, which lies on the principal axes. Therefore, any shape in the same class can be represented by the following equation:

$$x = \bar{x} + \Phi b$$ (2.1)

Here, $x$ is the target shape, $\bar{x}$ is the mean shape, $\Phi$ is a set of eigenvectors which represents modes of variations and $b$ is the shape parameter vector representing values for every mode. For a given $s$ shapes, the method of computing PCA is illustrated in [1] as follows:

1. Compute the mean vector

$$\bar{x} = \frac{1}{s} \sum_{i=0}^{s-1} x_i$$ (2.2)
CHAPTER 2. FACIAL FEATURE DETECTION AND TRACKING BACKGROUND

2. Compute the covariance matrix

\[ S = \frac{1}{s-1} \sum_{i=1}^{s} (x_i - \bar{x})(x_i - \bar{x})^T \]  \hspace{1cm} (2.3)

3. Compute a set of eigenvectors \( \Phi = \{\phi_0, ..., \phi_s\} \) from the covariance matrix \( S \) and sort them by their eigenvalues \( \Lambda = \{\lambda_0, ..., \lambda_s\} \) with \( \lambda_i \geq \lambda_{i+1} \) in descending order.

Figure 2.3 shows an example of the mean shape results from 968 shapes that were normalised and aligned with 68 two-dimensional (2D) points. Since every shape is represented by an \( nd \) dimensional vector, the shape vector is a 136-dimensional vector in the example above. PCA is then applied to compute the modes of variation ‘eigenvectors’ \( \Phi \). After applying the PCA, each group of landmarks can be represented in local coordinate space, and each shape can be presented by the linear equation, 2.1. Since the total value of the model variations is represented by the total sum of all the eigenvalues \( V_T = \sum \lambda_i \) [1], a proportion of the variations (e.g., 98%) found in the training set can be captured by selecting a subset of the eigenvectors where the sum of their eigenvalues represents 98% of the total variance. This process is conducted to avoid overfitting the model. The figure below presents an example of three variation modes, each of which varies between \( \pm 3\sqrt{\lambda_i} \) standard deviations.
Constraints must be applied to prevent the model from generating implausible shapes when using Equation 2.1. This can be done either by imposing limits on the derived distribution, such as \( p(b) \geq p_t \), where \( p_t \) is some threshold, or hard limits, such as the Mahalanobis distance as

\[
\left( \sum_{i=1}^{t} \frac{b_i^2}{\lambda_i} \right) \leq M_t
\]  

(2.4)

where \( M_t \) is a threshold chosen from the \( \chi^2 \) distribution. The thresholds can be chosen based on the desired criteria, such as 98% of the original data must pass the thresholds. This makes the \( b \) parameter legitimate, even when the distribution is not a Gaussian so that implausible shapes can be generated. However, Cootes and Taylor [1] suggested an effective approach for modelling the non-linear distribution of the parameter \( b \) using a mixture of Gaussians for kernel density estimation.

### 2.4 Fitting the Model to a New Image

In order to fit the generated instance of the model \( x \) with parameter \( b = 0 \) in a plausible shape \( x' \) where \( x' \) is the best approximate to the target the shape in the real image \( x' \), we must first project the shape into the parameter space by \( b' = \Phi^T(x' - \bar{x}) \), as suggested by [1]. Then, the optimal parameter \( b_i \) will be selected according to its distribution, such that \( p(b_i) \geq p_t \). If the distribution is a Gaussian, the threshold \( p_t \)
can be chosen, such as limiting it to three standard deviations $|b| \leq 3\sqrt{\lambda_i}$ or using the Mahalanobis distance, as in equation 2.4m and scaling $b_i$ until the threshold is met. However, if the distribution is not a single Gaussian, Cootes [1] describes a mix of Gaussian for kernel estimation to compute $p(b)$ and move it along the gradient ascent until a “local maxima” threshold is reached.

Before computing the vector of plausible shape parameters $b'$, it is assumed that the global transformation effects of scale $s$, rotation $\theta$ and translation $X_t, Y_t$ for shape $x'$ will be removed in order to align the shape with the model instance of the mean shape $x = \bar{x} + \Phi b$ in the reference frame. This will allow for further analysis of parameter $b'$, which will update the model instance to the best legitimate approximation. Therefore, the pose parameters $(T_s, \theta, X_t, Y_t)$ that best map $x'$ to the reference frame of the model instance $x$ must be computed. The procedure described by Cootes and Taylor [1] is as follows:

1. Generate an instance of the model $x = \bar{x} + \Phi b$.
2. Compute the pose parameters which map the instance $x$ to $x'$
3. Compute the inverse of the pose parameter and use it to transform $x'$ to the reference frame, $x'_\text{ref}$.
4. Scale $x'_\text{ref}$ by $1/x'_\text{ref}$ to generate $x_{tg}$, which aligns it to the tangent space of $x$.
5. Calculate the model parameter $b' = \Phi^T(x_{tg} - \bar{x})$ constrained respectively to the distribution of $b_i$.
6. If no significant change is found in either the pose or the shape parameters, the process converged; otherwise, return to 1.

### 2.5 Statistical Appearance Models SAMs

Appearance models have been used in many computer vision applications to interpret objects in scene images [19, 28, 29]. Face appearance models provides a foundation for many applications such as automatic recognition of identity and facial expression or
certain behaviour interpretation in general. These models can also be used in other application domains, such as face animation. Similar to SSMs, a statistical model of appearance developed by Cootes and Taylor [1] uses shape and texture variations found in the training set to build a model that can synthesise a new image. Along with SSM, it uses a statistical model of the texture variation that is found across the training set. Shape landmarks are used to outline the key features and as controlling points for the warping process. The aim of the warping technique is to use control points to align the variant feature locations with those in the mean shape. Therefore, SAMs can be built from sampled patches of image intensities in the \((g_{im})\) a shape-normalised frame. The warping technique in [1] uses a triangulation algorithm, namely the Delauney algorithm, which partitions an object image using a set of triangles and performs an affine transformation to map the triangles corners to their specified locations and moves all other inner pixels to their relative locations on the triangle. The model of appearance is similar to the eigenface model [30], but it is used in combination with SSM and computed in a shape-normalised frame.

In order to reduce the noise found in various textures, such as lighting, a sample \(g\) is generated from the warped image and normalised using a scale of \(\alpha\), which is the dot product of the warped patch with the mean image divided by the offset \(\beta\), which is essentially the average intensity or colour. The normalised patch is calculated by:

\[
g = \frac{(g_{im} - \beta1)}{\alpha}
\]  

The values of parameters \(\alpha\) and \(\beta\) are chosen to best match the normalised mean texture. This can be computed using the following equation [1]:

\[
\alpha = g_{im} \cdot \bar{g}, \quad \beta = (g_{im} \cdot 1)/n
\]  

where \(n\) is the number of vector elements.
CHAPTER 2. FACIAL FEATURE DETECTION AND TRACKING BACKGROUND

The mean texture is built in a similar way to the mean shape of SSM: by selecting an initial image to be the mean texture \( \bar{g} \) and iterating until it converges. The mean of \( \bar{g} \) is normalised so that it sums to zero with a unit variance. This process allows for a linear model to be built by performing another PCA analysis such that:

\[
g = \bar{g} + P_g b_g
\]  

(2.7)

where \( \bar{g} \) is the normalized mean texture vector, \( P_g \) is the vector of orthogonal modes of variations and \( b_g \) is the parameter value for every mode in \( P_g \). Equation 2.7 represents a linear model, which is used to obtain the original normalised texture that was effected by the scale \( \alpha \) and offset \( \beta \). Therefore, the following equation [1] allows the original warped patch to be retained:

\[
g_{im} = T_u(\bar{g} + P_g b_g) = (1 + u_1)(\bar{g} + P_g b_g) + u_2 1
\]  

(2.8)

where \( u_1 = (\alpha - 1) \) and \( u_2 = \beta \), so the first part of the equation \( (1 + u_1) \) returns the original scale while the last part \( u_2 1 \) adds the original offset.

To outline the steps of building the texture model, the algorithm can be summarized as follows:

1. Build a statistical shape model to obtain a mean shape \( \bar{x} \).
2. Warp all image examples to match their control points (shape landmarks) to the mean shape \( \bar{x} \).
3. Choose an example image to be the mean image texture \( \bar{g}_{im} \).
4. Normalize all images to best match the mean image by computing the appropriate scale and offset using equation 2.5.
5. Perform a PCA analysis for modes of texture variations to build the model of equation 2.7.

Since the model parameter \( b_g \) controls the change in the texture model and parameter
b_s controls the change in the shape model, a further PCA analysis on those parameters can yield a combined shape and texture parameter. However, to represent both texture and shape parameters b_g and b_s as one parameter, c, a weighting matrix W is applied to shape parameter b_s to account for the differences that allow both b_s and b_g to be mapped to a single parameter, c[1]. Usually, the weight matrix W = rI, where r is the ratio of the total variance in the texture model to that in the shape model. Then, PCA is applied so that the modes of variation can be presented as:

\[
b = \begin{pmatrix}
    W_s b_s \\
    b_g
\end{pmatrix}
= \begin{pmatrix}
    W_s P_s^T (x - \bar{x}) \\
    P_g^T (g - \bar{g})
\end{pmatrix}
\Rightarrow P_c c
\]  

(2.9)

where P_c represents the eigenvectors for both shape and texture, and c is the vector of values for each mode. P_c is represented as:

\[
P_c = \begin{pmatrix}
    P_{cs} \\
    P_{cg}
\end{pmatrix}
\]  

(2.10)

where P_{cs} is the shape eigenvectors and P_{cg} is the texture eigenvector. Both shape and texture can be modelled as in [1]:

\[
x = \bar{x} + P_s W_s^{-1} P_{cs} c , \quad g = \bar{g} + P_g P_{cg} c
\]  

(2.11)

which is summarized as:

\[
x = \bar{x} + Q_s c , \quad g = \bar{g} + Q_g c
\]  

(2.12)

Image synthesis is achieved by generating the shape instance and image g and then warping its control points to the shape x.

In order to match an unseen image marked with a set of landmarks x, the shape-texture parameter c can be computed after obtaining parameter b using equation 2.9, so that
c is given by:

\[ c = P_c^T b \]  

(2.13)

Then, \( c \) is applied to equation 2.11 to estimate the image patch and shape.

To evaluate the model fit to an image, a set of parameters, which can be used as ground truth data, must be estimated. Since the statistical model of appearance encompasses both texture and shape models, each requires a set of parameters suitable to the respected model. Cootes and Taylor [1] developed popular approaches to perform this measurement, including the \textit{active shape model (ASM)} and \textit{active appearance model (AAM)}. The following sections provide brief overviews of these models.

## 2.6 Active Shape Models (ASMs)

Kass et al. [31] presented a study on deformable models that used flexible contour tracking \textit{“snakes”} for every feature in order to deform the shape models to better fit the image. The deformation of the model relies on a snakes local search to predict the required displacement for each model point. Another ASM called \textit{“smart snakes”}[32] applies global shape constraints derived from statistical shape models to restrict the deformations so remain consistent with the shapes learned from the training set. Global constraints are built based on the distribution type in the PDM, such as the Mahalanobis distance.

The ASM algorithm can be used to perform a local search to optimise the fit of the shape \( x \) to an image, as well as to displace \( x \) to \( x' \) by estimating the appropriate shape and posing parameters \( b \) and \( T(s, \theta, X_t, Y_t) \). The algorithm estimates the displacements for each point between \( x_i \) and \( x'_i \) by examining the region around \( x_i \) and computing the required parameters in order to deform the shape \( x \) legitimately and best match \( x' \). The regions around each point are examined to find the best match based on the statistical model of the texture that is observed in the training set. A local search and
fit are carried out iteratively until no significant changes are found. Briefly, the ASM steps are as follows [1]:

1. Examine a region of the image around each point $x_i$ to find the best nearby match for the point $x'$.  
2. Update parameters $(X_t, Y_t, s, \theta, b)$ to best fit the newly found point $X$.  
3. Repeat until convergence occurs.

### 2.6.1 Local Models for Each Point

The quality-fit measure in [1] relies on statistical models of samples that are built from the training set around each model point, thus allowing measures of similarity to be used between the test sample and those derived from the training set. The following sections provide an overview of how these local models are constructed, as well as their evaluation methods.

The model for every point $x_i$ across the training set is built using sampling profiles that are normal to the model boundary of length $k$ on both sides of point $x_i$ so that the length of the profile is $2K + 1$. Derivatives of these samples that resemble vector $g_i$ are then used and normalised by dividing each vector element value by the sum of the absolute values in the sample. This is given as [1]:

\[
g_i \rightarrow \frac{1}{\sum |g_{ij}|} g_i
\]  

Since the distribution of $x_i$ is assumed to be a multivariate Gaussian, the models’ parameters of profiles can be linearly expressed with the Mahalanobis distance, which can be computed as follows [1]:

\[
f(g_s) = (g_s - \bar{g})^T S^{-1} (g_s - \bar{g})
\]
The function \( f(\mathbf{g}_s) \) gives the distance of the point from the mean, thus allowing the optimal fit to be determined by searching for the minimum value of \( f(\mathbf{g}_s) \). This maximises the probability that the sample will fall within the distribution.

When searching for a match to the sample profile \( 2k + 1 \), the sample slides in a longer profile that is normal to the boundary defined by \( 2m + 1 \) where \( m > k \), as shown in Figure 2.5.

![Figure 2.5: Profile search along normal to model boundary; the red dot shows the current profile position in the search.](image)

This gives the test sample \( m - k \) possible locations, as indicated by the dots in the figure above, which are evaluated by recording the cost of \( f(\mathbf{g}_s) \) at the given position. After all of the positions are examined, the position that best minimises the cost is selected.

Cootes and Taylor [1] presented extensions to improve robustness against noise, such as occlusion and clutter, while Burt [33] suggests using image pyramids to build statistical models that are based on various resolutions. The search is performed with multiple image resolutions starting from coarse and moving to a finer level, thus refining the point location on multiple levels. The process iterates for all points in the model until sufficient points are found. The final result is given at the highest resolution (Level 0) in the image pyramid. In [1], a three-pixel width for parameter \( k \) is used to suit various application domains.
2.7 Active Appearance Models AAMs

Unlike ASMs, which use limited texture information on the target object to locate a set of points, AAMs [19] take into consideration all texture information on extracted patches. Using an initial estimate of the object location, the AAM provides a way to synthesise a new image that approximates the real image using a statistical model of appearance, as described above. The AAM aims to understand the relationship between the spatial patterns found in $\delta I = I_i - I_m$, where $I_m$ is the appearance synthesised by the model and $I_i$ is the target image under the model, and the residuals of the model parameters $\delta p$ across the training set. The model then applies this relationship to an iterative search in order to find the model parameter that best matches the image.

With the new image, the AAM aims to predict suitable model parameters that will best minimise the magnitude of $|I_i - I_m|^2$. Since the AAM consists of many parameters, this can be considered a difficult optimisation problem. Instead of examining all of the model parameters to find the one that best minimises $\delta I$ (i.e., the difference between synthesised and real images), which can be very difficult and computationally expensive, Cootes [19] has presented an efficient approach that is described briefly below.

As discussed above, the parameter $c$ controls shape and texture models (equation 2.12) where $Q_s$ and $Q_g$ describe the modes of variation. The instances of the models are first generated in the model frame in the form of $A : A = T_i(a)$ where $A$ represents the instance of the model in general, which can be a shape $X$ or texture $G$, and $T_i$ represents the required transformation parameters of the model, which is a similarity transform for the shape model $T_i(x) = T(x; s, \theta, x_t, y_t)$, and a normalisation for the appearance model $T_u(g) = T(g; \alpha, \beta)$. The transformation, $T(a)$, initially represents the identity transformation of the model. The transformations are later represented as $T_i(T_{i\delta t}(a)) = T_{i+\delta t}(a)$.
First, the shape model instance is generated in the form of $\mathbf{x} = \mathbf{x} + \mathbf{Q}_s \mathbf{c}$ by applying the appropriate pose $\mathbf{t}$ and model parameter $\mathbf{c}$. Then, the appearance model instance $\mathbf{g} = \mathbf{g} + \mathbf{Q}_g \mathbf{c}$ is generated. Given a generated shape $\mathbf{x}$, the appearance model instance image is warped to match its control points to the shape points. This is done in a shape-normalised frame, as described above. For the given AAM parameters $\mathbf{p} = (\mathbf{c}^T | \mathbf{t}^T | \mathbf{u}^T)$, the current residual can be measured as:

$$
\mathbf{r}(\mathbf{p}) = \mathbf{g}_s - \mathbf{g}_m \quad (2.16)
$$

For a given residual vector $\mathbf{r}$, the scalar error measure can be given by the sum square error $\mathbf{E}(\mathbf{p}) = \mathbf{r}^T \mathbf{r}$. The aim is to estimate the parameter change $\delta \mathbf{p}$, which minimises $\delta \mathbf{I}$ using the first order of the Taylor expansion:

$$
\mathbf{r}(\mathbf{p} + \delta \mathbf{p}) = \mathbf{r}(\mathbf{p}) + \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \delta \mathbf{p} \quad (2.17)
$$

where the $ij^{th}$ element of matrix $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ is $\frac{\partial r_i}{\partial p_j}$. By equating 2.17 to zero, the RMS can be used to estimate $\delta \mathbf{p}$, which is given by:

$$
\delta \mathbf{p} = -\mathbf{R} \mathbf{r}(\mathbf{p}) \quad \text{where} \quad \mathbf{R} = \left( \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \right)^{-1} \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \frac{\partial \mathbf{r}^T}{\partial \mathbf{p}} \quad (2.18)
$$

Since measuring $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ in the shape-normalised frame yields minor variations, it can be considered approximately static. Therefore, one can estimate $\frac{\partial \mathbf{r}}{\partial \mathbf{p}}$ via numeric differentiation by automatically changing each element of $\mathbf{p}$ from its optimal value on either typical or synthesised images and then averaging them over the training set. The degree to which displacement is limited to $\mathbf{p}$ depends on the type of the model, which is determined experimentally. Cootes and Taylor [1] found that a half standard deviation in relative distribution for every element in $\delta \mathbf{p}$ adequately represented large variations in residuals in face models. A Gaussian kernel is added for smoothing, and the full
The constrained local model is given by:

\[
\frac{dr_i}{dp_i} = \sum_k w(\delta c_{jk})(r_i(p + \delta c_{jk}) - r_i(p)) \tag{2.19}
\]

where \(w(x)\) is a proper normalized Gaussian kernel of weights. Then, \(R\) is estimated again and used to estimate the required change in model parameters in \(\delta p\) for new searches that use equation 2.18.

Given a matrix \(R\), the required change for each parameter \(\delta c_i\) is given by

\[
\delta c_i = rw_i \cdot \delta g \tag{2.20}
\]

where \(rw_i\) represent the \(i^{th}\) row of \(R\), and the \(\delta g\) is the vector of differences between the model intensities and underlying image intensities after projection into the model frame.

The predicted change in \(\delta c\) is further optimised in the iterative method described by Cootes and Taylor [1], which is applied in a multi-resolution scheme that measures the error vector after each iteration. It also selects the change of \(\delta c\), which would reduce the error in \(\delta g\). However, if no improvements are found, different scales of parameter \(\delta c\) are also examined. Nevertheless, it is essential to note that since the shape model is combined with the AAM, changes made by the AAM to reduce error vector \(\delta g\) would affect the shape model. The next section describes a more sophisticated model called the *Constrained Local Model*.

### 2.8 The Constrained Local Model

The idea of building a deformable model that can utilise shape and texture has attracted many researchers [19, 24, 34–36]. Identifying the correlation between feature locations allows the shape model to accurately predict partially occluded models to
some extent. The following section provides an overview of several models that learn variations in shape and texture:

The *Pictorial Structure Matching (PSM)* algorithm [36] uses a set of texture templates for each model part. In the human case, the arm, leg, head and so on are built into a tree structure. The range of relative distances between each pair of limbs is also learned from the training set. PSM optimisation is achieved by an objective function that maximises the match in both templates’ match responses and the distances between each pair of limbs. This method can be used to conduct a global object search; however, instead of using a complete shape model, correlations of pair-part locations are modelled into a tree structure.

The AAM approach uses a combined linear model of texture and shape to generate image patches using an initial estimate of the shape model. For example, in face models, the AAM uses a large texture vector to represent the face patch, which is sampled at each iteration. This can reduce algorithm inefficiency by eliminating expensive computations, especially when a large number of pixels is used to represent face patches.

Both ASM and AAM require a good initial approximation of the shape model location and are sensitive to local minima. However, the local multi-scale Gaussian derivatives as used in [37] to conduct a feature point search can also be applied to overcome this problem, as described in [35].

Cristinacce and Cootes [38] presents a facial feature tracker that uses the AdaBoost template with Haar wavelet-like features for each local feature. The algorithm employs the efficient face detector described by Viola and Jones [39] to narrow the scope of the search to the face object. Then, local templates that search for local features are trained to classify false and positive samples using a data structure for every feature that is built from a set of Haar wavelet-like responses. False samples are chosen by identifying small random displacements from the true location. The search scope is
narrowed as the global face detector and shape model provide an initial approximation of the face and feature locations. This technique evaluates the candidate regions using a probability density function that is learned from the distribution of histograms of feature responses of both negative and positive samples. During the search, the shape is fit to the current estimate, and the plausibility of the shape is evaluated using the constraints on the orientation, scale and location of the features, relative to those of the detected face region.

[35] introduced a similar technique called *shape-optimised search (SOS)*, which estimates the face position using Viola and Jones’ face detector [39]. The position variations in local features within the face region are learned from the training set. Each face feature is applied to its range of estimates, such as scale and orientation, relative to the face region. The response is computed to individual features in order to form the *feature response surface* using an independent template-based detector. Three detectors, namely normalised correlation, orientation maps[40] and boosted classifiers [39] are shown, which is similar to the work of X. The responses are then evaluated in the feature’s relative region and the average response is chosen relative to each feature point for the initial estimate of the shape model. Then, a constrained shape model is fitted to these points using a non-linear optimiser called the *Nelder-Mead simplex method*[41] in order to find the best fit.

Following the work of Cristinacce and Cootes [35], Cristinacce and Cootes [34] presented a *CLM* that uses the global face detector described in [39], and similarly, a set of local feature detectors within the face region where their positions are approximated relative to the detected face using PSM constraints [36]. This creates a good initial estimate for the shape model. However, instead of using fixed templates for all of the iterations, as in [35], a set of dynamic feature templates are updated using the nearest neighbour method to select the appropriate example from the training set. The process of generating a template model in [34] is similar to that of the AAM model [19]. However, instead of using one large template to model the face patch variation, smaller
templates are used to model the variation around each local feature. The texture grey level values of the patches in a given image are concatenated in a single grey level, $g$, and the steps of building the joint texture-shape model proceeds in a similar way to that of the AAM. For the search, the objective function used is the sum of the template responses and the log-likelihood of the model parameters $p$:

$$p = (t^T | b_s^T)^T$$

where $b_s$ is the shape parameter and $t$ is the transformation parameter scale, orientation and translation. The objective function is given by:

$$f (p) = \sum_{i=1}^{n} I_i (X_i, Y_i) + K \sum_{j=1}^{s} \frac{b_j^2}{\lambda_j}$$

(2.21)

where $n$ is the number of features, $I_i (X_i, Y_i)$ is the response of a given patch and $K$ is the ratio of $\sum_{i=1}^{n} I_i (X_i, Y_i)$ and $\sum_{j=1}^{s} b_j^2 \lambda_j$ is verified using a test set. The weighting $K$ is similar to the PSM constraints that are found in the pair parts, which are used to describe the Mahalanobis distances. Therefore, $K$ is applied to represent the level of plausibility of the shape parameters. Parameter $b_j$ is the distance of the feature point in the model frame and $\lambda_j$ is the variance of $b$. $\sum_{j=1}^{s} \frac{b_j^2}{\lambda_j}$ indicates the sum of distances in standard deviation units for the shape points. Then, the Nelder-Mead simplex [41] method is used to find the optimum parameters.

Cristinacce and Cootes [34] proposed an algorithm for static images, as well as an efficient facial-feature tracking mechanism, using the a template selection tracker (TST), as described in [42]. In this context, it is able to select the best template based on the nearest neighbour method. Later [43] describes this research in more detail, and the algorithm steps can be summarised as follows:

1. Estimate the initial input.
2. Fit the joint model of the shape and texture to the current estimates.
2.9. **DECISION TREES**

3. Generate new templates using the current estimates.
4. Find the best fit for the new shape using current template responses.
5. Terminate when convergence occurs, otherwise return to 2.

If no significant change in shape parameters and templates are found, then convergence is declared, and the process terminates.

### 2.9 Decision Trees

The concept of decision trees has existed for a long time [44], and they have been adopted widely. Decision trees' have become popular because generalisation allows a predictor that is sufficiently trained on a set of examples to generalise its decision to any other unseen examples. Each decision tree consists of a set of nodes and leaves. In binary trees, the starting node is referred to as the root node, which splits into two nodes. Subsequent nodes are created in the same hierarchical fashion. Each node contains a binary split function that splits the data into two different subsets. When the tree grows, the amount of data that reaches the descendant node decreases. When the splitting reaches its minimum (via a threshold), the splitting function is terminated and the last node, which holds the data, is called a leaf. The splitting function is optimised by a measure called entropy, which measures the level of separability and allows the node to determine the best feature for the split function.

### 2.10 Boosting

Boosting-based algorithms are widely used in machine learning and computer vision research [45–47]. In ensemble learning, training a predictor with sufficient training data yields a single complex strong learner. On the other hand, with weak learners, each individual learner is trained using a subset of the data. Boosting is used in ensemble learning that combines multiple weak learner models to yield a model that outperforms
a single strong learner. The probability of errors for each individual weak learner is significantly higher than that of a strong learner, which is usually slightly less than 0.5 [44], but not when they are combined. Building a set of weak learners is much easier and simpler than using a strong learner in complex cases. There are many ways to combine weak learners’ estimates after they have been weighted, such as by majority or average. The complex combination method can be used for the majority if all weak learners have equal weights. For instance, AdaBoost, which stands for Adaptive Boosting, uses a sequence of predictors that are trained on a sequence of versions of weak learners’ estimates. Each predictor in the main sequence is trained to reduce the error from the result that is produced by the previous predictor.

2.11 Random Forest

Breiman introduced the idea of Bootstrap aggregation 'bagging' in [48]. This technique is usually used to generate a large number of drawn samples with replacement from a specific dataset. This is especially useful when large numbers of samples are needed to reduce the variance of decision trees and improve decision stability via the aggregation of bootstrap replicates. Later, Breiman [49] introduced Random Forests (RFs), which combine the ideas of bagging and random feature selection. This concept was introduced independently in [50–52]. RFs are appealing to many researchers because they provided multiple features, such as unsupervised learning, differential class weighting to deal with different sample sizes and representing missing data [44]. RFs can be applied to classification and regression tasks using either categorical or continuous random variables. Another advantage of using RFs is that they can measure the importance of variables. Since many algorithms apply PCA to reduce data dimensionality before fitting a predictor, the information loss caused by PCA can be vital to maximising the prediction potential. [44] recommends evaluating the variables’ importance and applying fitting only to important predictors, which is a better approach.
The computer vision community has recently begun to realise the benefits of RFs and have applied them to many algorithms [22, 46, 53]. Cootes et al. [22] recently developed an algorithm that has shown an impressive ability to locate feature points. The method uses CLM [43] to constrain the shape and RF-based voting as a feature template where each decision tree votes for the optimal feature location. The following section describes this algorithm in more detail.

2.12 Combined CLMs and Random Forests

Cootes et al. [22] presented a regression-based voting scheme that uses RFs to generate high-quality cost map images to describe the degree of confidence at every point, \( z \), in the region of interest. The predicted points are estimated relative to \( z \) using a set of features \( f(z) \). This is achieved by using a forest of randomised regressors where each regressor produces a weighted vote that is added to the accumulator array, \( V \). Each regressor is trained to predict the best position of a point using a set of features that is chosen randomly from all possible Haar wavelet-like features. Bootstrap with replacement is used as the sampling method for each tree. The accumulator array is then smoothed to allow for uncertainty in the predictions. This creates a cost-map image where a shape model is fitted using the CLM [43]. The CLM can be used to regularise the output of the predictors because there are various ways to combine regressor votes, as described by [22]. However, single voting is shown to have the best performance.

During the training, features are sampled in a reference frame where the samples are displaced by a small distance from the true position. This enables the tree to learn the spatial patterns relative to the displacement of parameters \( p \) and \( t \). Each tree then learns the feature and threshold that will best split the two classes of samples using a simple entropy measure. The objective is to minimise the entropy of each class.

When searching in an unseen image, a global face detector [39] is used to locate the face
region. This process is repeated for every local feature. The global detector returns a rectangular area for each candidate face, which is used to initialise the SSM model. After initialising the points to the current estimate, the CLM performs a finer search and optimises the quality of the fit using the cost images produced by regressors \( C(x_i) \), as shown by:

\[
Q_0(b, t) = \sum_{i=1}^{N} C(x_i) \quad \text{Subject to} \quad b^T S^{-1} b \leq M_t \tag{2.22}
\]

This filters out implausible shape parameters, which are created by \( b \) and are rejected using threshold \( M_t \), which is chosen using CDF in the \( \chi^2 \) distribution.

### 2.13 Summary

Facial feature detection is a challenging task and is considered as pre-requisite for many facial expression analyses. Despite that facial detection and tracking are a fast moving fields, tracking under uncontrolled settings remains challenging. There has been a significant shift towards CNN-based method. This is due to their success in multiple computer vision tasks. However, a key advantage of Random Forest Regression Voting Constrained Local Model (RFCLM) is that they provide computationally efficient and accurate detections. In this chapter we reviewed several approaches that contribute to the development of RFCLM. While RFCLM can be used to build a robust generic facial tracker, subject-specific AAM models can provide superior performances.

There are several configuration parameters associated with RFCLM and AAM trackers. In the next chapter we perform extensive experiments in order to optimise these configurations.
Chapter 3

Facial Tracking Selection and Optimisation

3.1 Introduction

Automatic facial behaviour analysis involves automatic facial feature tracking. Thus, ensuring the reliability of tracked data is crucial to the analysis task, and we optimise our facial feature tracker to increase data reliability while maintaining adequate efficiency in real-time applications. To optimise our facial feature tracking we performed the following:

- Using a multi-stage search strategy where, in the initial stages, the detection is limited to a few stable features using low-resolution images (e.g., eye and corners of the mouth). Then, the results of initial models can be used to guide the local search towards finer resolution images, which can include a greater number of features. This reduces the scope of the search and improves the initialisation of consequent models, increasing the accuracy and efficiency of the algorithm.
• Include two types of facial tracking systems, RFCLM and AAM. The RFCLM-based tracker is used as the main generic tracker, while AAM is used as a person-specific tracker. AAM is only used with subjects that exhibit extremely challenging conditions.

• Optimise the main configuration parameters associated with training the aforementioned tracking techniques using several initialisation scenarios. This is achieved via an extensive experiment conducted to identify the optimal configuration parameters at each stage for each tracker.

• Evaluate the RFCLM parameters that best fit the so-called in-the-wild scenario. We exclude the assumption of good initialisation and instead use an out-of-the-box global face detector developed by the department to identify the optimal configuration parameters associated with the global detector.

• We also further evaluate the RFCLM when used with a robust global face detector (built locally by the department) to track facial features on video sequences of spontaneous facial expressions using the public dataset DISFA[2].

3.2 The dataset

Since AAM is used as a non-generic tracker in this project, only RFCLM-based test results can be validated on public datasets, while AAM needs to be tested on relevant subject samples. Several public datasets are used to train and test RFCLM: labelled face parts in-the-wild (LFPW)[54], annotated Faces in-the-wild (AFW)[55], Helen[56], multi-modal verification for teleservices and security applications (XM2VTS)[57], and intelligent behaviour understanding group (IBUG)[58]. These datasets’ major characteristics are as follows.

• **Labelled face parts in-the-wild (LFPW)[54]**: Due to copyright issues, this dataset only provides the URLs of 1,432 facial images collected from various website sources, such as Flicker, Google, and Yahoo. The dataset contains facial
images captured in naturalistic settings with highly variable poses, facial expressions, lighting conditions, and occlusions. However, some of these URLs are outdated and no longer accessible. Ground truth annotations are provided for the whole dataset and describe the position of 35 facial landmarks. The annotation process was performed by three workers employed through Amazon Mechanical Turk (MTurk)[59], and the average position is used as the ground-truth.

- **Annotated faces in-the-wild (AFW)[55]**: This dataset consists of 205 images with 486 faces in naturalistic settings. The images were collected from Flicker.com and annotated with six facial landmarks.

- **Helen[56]**: This dataset contains 2,330 facial images collected from Flicker.com using several search keywords, such as portrait, family, and outdoor. The collection method involves using several languages to ensure culturally unbiased samples. Most of the facial images are high resolution with a face width greater than 500 pixels. This dataset is also rich in terms of available annotations. It contains 194 facial landmarks initially performed by workers employed through MTurk[59] and later reviewed and verified by the authors.

- **Multi Modal Verification for Teleservices and Security Applications (XM2VTS)[57]**: This dataset contains images of 295 subjects captured at four time-points. These images are of high quality with subjects in frontal poses under constrained conditions. A 68-landmark annotation of this dataset is available from [60].

- **Intelligent behaviour understanding group (IBUG)[58]**: This dataset contains 135 naturalistic facial images annotated with 68 facial landmarks. Images in this dataset display faces with challenging poses involving the head and facial expressions.

Original landmark annotations for these datasets differ in both quantity and semantics (see Table 3.2). However, all these datasets were re-annotated under one scheme in *The 300 Faces In-the-Wild Challenge* [58] following the Multi-PIE[61] annotation scheme,
which enables learning from larger sample variation. We use all of the above-mentioned datasets to train and evaluate the RFCLM. There are two versions of the 300W test datasets. The first test dataset consist of three subsets: the test datasets of HELEN and LFPW, as well as the IBUG. It is essential to note that under 300W challenge, the HELEN and LFPW test subsets are commonly labelled as the common dataset, while the IBUG dataset is labelled as the challenging dataset. The second version of 300W was originally a private dataset used for the competition. This dataset consists of 300 indoor and outdoor facial images, 600 in total. We also used this dataset to evaluate our optimised RFCLM.
### The Parameters and Scope

The detection of facial features in RFCLM is performed by a number of local experts corresponding to the number of facial features. Each local expert is trained on samples around the relevant feature. The training samples are defined by several parameters in RFCLM training, such as the reference frame width, patch size, and maximum random displacement distance in $x$ and $y$ directions.

The reference frame defines a uniform scale for all facial images using a reference length; in this case, we use the external corners of the eyes as the reference length (see Figure 3.1). The patch defines the size of a rectangular area inside the reference frame from which samples are drawn and used for feature generation and learning. The maximum displacement distance defines the maximum distance used to displace the patch inside the reference frame. There are also other parameters, such as a fixed search border and the proportion of noise added to the scale and orientation of sampled frames. These settings are left fixed in our optimisation experiments.

---

### Table 3.2: Several public datasets for facial images (T: training dataset, E: Evaluation dataset)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Original samples</th>
<th>annotations</th>
<th>re-annotated samples by [58]</th>
</tr>
</thead>
<tbody>
<tr>
<td>HELEN[56]</td>
<td>2,330 (T=2000, E=330)</td>
<td>194</td>
<td>2,330 (T=2,000, E=330)</td>
</tr>
<tr>
<td>AFW[55]</td>
<td>205 (486 faces)</td>
<td>6</td>
<td>185 (337 faces)</td>
</tr>
<tr>
<td>IBUG[62]</td>
<td>135</td>
<td>68</td>
<td>135</td>
</tr>
<tr>
<td>LFPW[54]</td>
<td>1432 (T=1132, E=300)</td>
<td>35</td>
<td>1035,(T=811,E=224)</td>
</tr>
<tr>
<td>XM2VTS[57]</td>
<td>2,360</td>
<td>68</td>
<td>2,360</td>
</tr>
<tr>
<td>300W [58]</td>
<td>689</td>
<td>68</td>
<td>689</td>
</tr>
<tr>
<td>300W-2 [58]</td>
<td>2x300 (indoor+Outdoor)</td>
<td>68</td>
<td>600</td>
</tr>
</tbody>
</table>
CHAPTER 3. FACIAL TRACKING SELECTION AND OPTIMISATION

Figure 3.1: Distance error is normalised by the outer-ocular distance only when optimising RFCLM. Final evaluations on 300W datasets were normalised by the inter-ocular distance for comparable results.

Figure 3.2: Selected RFCLM parameters to optimise

We examine the effects of varying RFCLM parameters under various initialisation assumptions, from low-to-high distance error from the ground-truth position. To reduce the computational time, we only consider varying the main parameters, which are the reference frame width, patch size, maximum distance of displacements, and a binary scaling for the search range, which acts as a switch for activating the displacement beyond the search border. The search is conducted in a rectangular region around the point of the half-width \( h_s = s \times p_r + \text{border} \), where \( p_r \) is the half-width of the bounding box containing the training examples of the point, and a border is added to account for uncertainties in the initial position. (see Figure 3.2).
3.4 EXPERIMENT DESIGN

We design our experiment to systematically examine a varying range of possible values of the aforementioned parameters in a multi-stage and multi-level manner as shown in Figure 3.3. In the first stage, we focus on detecting a small subset of salient features, such as the corners of the mouth and eyes. In the second stage, we use the results of the first stage to initialise the full 68-point model.

For each level in each stage, we examine the list of values shown in Table 3.3. We use all valid combinations of these parameters (156 in total). For example, we exclude combinations that specify a patch size greater than the frame size.

Table 3.3: Examined values of various parameters for 6 and 68 points RFCLM

<table>
<thead>
<tr>
<th>No</th>
<th>Frame Size</th>
<th>Patch Size</th>
<th>Random Displacements</th>
<th>search scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21</td>
<td>15</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>17</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>19</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>21</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>161</td>
<td>23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Since our optimisation involves hundreds of experiments, each exploring different combinations of RFCLM’s parameters, the training time can become an issue. To reduce the training time, we only select a subset of all the available data (see Table 3.2). We select the Helen dataset as it more culturally balanced with images captured in naturalistic settings, and faces are mostly frontally posed, which is suitable for the purpose of this project. It also contains a separate datasets dedicated for training and testing, which makes our evaluation results comparable with a public dataset. We use these subsets as the basis for our RFCLM parameter optimisation experiments.

### 3.4.1 Evaluation Criteria

Since RFCLM is dependent on initialisation, we evaluate each model while assuming various initial distance errors. We perform a random perturbation of the ground-truth points by a certain distance to simulate the effects of the error. We use a total of four maximum error distances, which are 30%, 20%, and 10% of the reference length and another 10 pixels in both $x$ and $y$ directions.

To evaluate the detection results of each model, we use a common error measure to evaluate facial feature detection accuracy by measuring the error distance relative to a reference length (i.e., the distance between two points) instead of pixels, which reflects the size of the detection error relative to the face size. We compute the error of each sample as the mean error recorded for all points normalised by the reference distance as shown in Figure 3.1 and Equation 3.1.

$$\text{error} = \frac{1}{n \times d_{\text{ref}}} \sum_{i=1}^{n} |\hat{x}_i - x_i|$$

where $\hat{x}_i$ is the ground truth positions of the features and $d_{\text{ref}}$ is the reference distance between the external eye corners.

We rank all the models based on six statistical measures as shown in Table 3.4, which are based on the cumulative distribution function (CDF) of the error measure described...
above to compare and select the final model for each level.

3.5 RFCLM optimisation

3.5.1 First stage

In the first stage, we aim to find the optimal sequence of models that produce the best and most robust detection results for the six facial features as shown in Figure 3.3. We plot the results of the top models (see Figure 3.4) for the first level in the first stage. The results demonstrate that for large initial errors, models with lower reference frame widths produce good initial results as shown in Figure 3.4a, while a frame width of 81 shows a better overall performance in all initialisation scenarios. For more details, we show some statistical measures of the error in Table 3.4.
CHAPTER 3. FACIAL TRACKING SELECTION AND OPTIMISATION

(a) 30% of reference length

(b) 20% of reference length

(c) 10% of reference length

(d) 10 pixels

Figure 3.4: Top 6-points RFCLM models evaluated under various initialised errors. Each curve represents an RFCLM with different values for the main configuration parameters. The reference frame width denoted by fw, the patch width denoted by pw, the displacement distance used in training RFCLM denoted by di and the search scaling denoted by s. Notice in the 30% displacement, lower frame size (fw=41) gives better initial estimates whereas in the rest, a frame width of 81 shows better overall performance.
Table 3.4: Optimal RFCLM for First Level in stage 1 (6 points). T-D: refer to training displacement distance while I-E refers to Initialised Error Distance either relative to the reference distance (%) or in pixels (px)

<table>
<thead>
<tr>
<th>No</th>
<th>RFCLM configurations parameters</th>
<th>Error Stat</th>
<th>CDF percentiles</th>
<th>I-E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frame width</td>
<td>Patch size</td>
<td>T-D</td>
<td>Scale factor</td>
</tr>
<tr>
<td>1</td>
<td>161</td>
<td>21</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>161</td>
<td>23</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>81</td>
<td>15</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>23</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>161</td>
<td>21</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>161</td>
<td>23</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>81</td>
<td>15</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>81</td>
<td>23</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>41</td>
<td>17</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>81</td>
<td>21</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>81</td>
<td>23</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>81</td>
<td>23</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>41</td>
<td>15</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>81</td>
<td>21</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>81</td>
<td>21</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>81</td>
<td>23</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>81</td>
<td>23</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>
We also examine whether the resulting top models are sensitive to different initialisations performed by different RFCLM models. For example, when used as a second level in the first stage (6-points search), we notice the selected model (row 14 from Table 3.4) still dominates all other results and shows robustness in all scenarios as shown in Figure 3.5.

Figure 3.5: Top results for a sequence of two RFCLMs searchers, with the second model fixed. The results suggest that the second model results are robust and only a small margin is influenced by the first model detection.
Table 3.5: Evaluation results when using the first RFCLM model as a second level. This table shows top experiment results and how RFCLM are robust to different initialisation produced from other models. T-D: refer to training displacement distance while I-E refers to Initialised Error Distance either relative to the reference distance (%) or in pixels (px).

<table>
<thead>
<tr>
<th>No</th>
<th>Frame width</th>
<th>Patch size</th>
<th>T-D</th>
<th>Scale factor</th>
<th>Error Stat</th>
<th>CDF percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90%</td>
</tr>
<tr>
<td>1</td>
<td>161</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>2.71</td>
<td>0.0212</td>
</tr>
<tr>
<td>2</td>
<td>161</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>2.71</td>
<td>0.0212</td>
</tr>
<tr>
<td>3</td>
<td>161</td>
<td>15</td>
<td>15</td>
<td>1</td>
<td>2.99</td>
<td>0.0333</td>
</tr>
<tr>
<td>4</td>
<td>161</td>
<td>15</td>
<td>7</td>
<td>1</td>
<td>2.7</td>
<td>0.0238</td>
</tr>
<tr>
<td>5</td>
<td>161</td>
<td>15</td>
<td>7</td>
<td>1</td>
<td>2.7</td>
<td>0.0238</td>
</tr>
<tr>
<td>6</td>
<td>161</td>
<td>23</td>
<td>11</td>
<td>0</td>
<td>2.92</td>
<td>0.0299</td>
</tr>
<tr>
<td>7</td>
<td>161</td>
<td>23</td>
<td>19</td>
<td>0</td>
<td>2.92</td>
<td>0.03</td>
</tr>
<tr>
<td>8</td>
<td>31</td>
<td>21</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>0.0331</td>
</tr>
<tr>
<td>9</td>
<td>31</td>
<td>23</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>0.0332</td>
</tr>
<tr>
<td>10</td>
<td>41</td>
<td>15</td>
<td>19</td>
<td>1</td>
<td>2.98</td>
<td>0.0318</td>
</tr>
<tr>
<td>11</td>
<td>41</td>
<td>17</td>
<td>19</td>
<td>1</td>
<td>2.97</td>
<td>0.0319</td>
</tr>
<tr>
<td>12</td>
<td>81</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>2.73</td>
<td>0.0211</td>
</tr>
<tr>
<td>13</td>
<td>81</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>2.73</td>
<td>0.0211</td>
</tr>
<tr>
<td>14</td>
<td>81</td>
<td>15</td>
<td>15</td>
<td>1</td>
<td>3.07</td>
<td>0.0373</td>
</tr>
<tr>
<td>15</td>
<td>81</td>
<td>15</td>
<td>7</td>
<td>0</td>
<td>2.68</td>
<td>0.0211</td>
</tr>
<tr>
<td>16</td>
<td>81</td>
<td>15</td>
<td>7</td>
<td>0</td>
<td>2.68</td>
<td>0.0211</td>
</tr>
<tr>
<td>17</td>
<td>81</td>
<td>15</td>
<td>7</td>
<td>0</td>
<td>2.91</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Given the best initial estimates by the first RFCLM, we further examine whether these results can be improved by using other models. As shown in Figure 3.6, the results demonstrate improvements but only on larger initialised errors.

Figure 3.6: Top 6-points RFCLM models evaluated under various initialised errors. Each curve represents an RFCLM with different values for the main configuration parameters. The reference frame width denoted by fw, the patch width denoted by pw, the displacement distance used in training RFCLM denoted by di and the search scaling denoted by s.
Table 3.6: Effects of using a second level. This table shows top experiment results and how top first RFCLM model performance is slightly increased only on low initialised errors of (10px and 10%) when using a larger reference frame width.

<table>
<thead>
<tr>
<th>No</th>
<th>RFCLM configurations parameters</th>
<th>Error Stat</th>
<th>CDF percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frame width</td>
<td>Patch size</td>
<td>T-D</td>
</tr>
<tr>
<td>1</td>
<td>161</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>161</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>161</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>161</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>161</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>161</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>161</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>161</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>161</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>41</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>81</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>161</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>41</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>81</td>
<td>23</td>
<td>11</td>
</tr>
</tbody>
</table>
3.5.2 Second stage

Similarly, another experiment is undertaken to evaluate RFCLM’s detection of all 68 points. All possible combinations of these parameters are tested. Figure 3.7a shows the top-ranked RFCLM in all four initialisation scenarios, and Table 3.7 shows the statistical results.

The top overall performance (e.g., at the 99th percentile) is shown with a reference frame width of 81 and patch size of 15 with a maximum displacement of 11. This is only true if the initialised error (I-E) is very small, within 10 pixels. For a larger I-E (e.g., 30r), a reference frame width of 41 with a patch size of 15 and maximum displacement of 15 shows the optimal overall performance.
3.5. RFCLM OPTIMISATION

Figure 3.7: Top 68-points RFCLM models evaluated under various initialised errors. Each curve represents RFCLM with a set of configuration where fw is the reference frame width, pw is the patch width, di is the displacement used in training RFCLM and s is the search scaling. The results suggest that smaller reference frames give better overall performance for larger initialised error (see Table 3.7)
Table 3.7: Best Performance of RFCLM for First Level in stage 2 (68 points). T-D: refer to training displacement distance while I-E refers to Initialised Error Distance either relative to the reference distance or in pixels (px).

<table>
<thead>
<tr>
<th>No</th>
<th>RFCLM configurations parameters</th>
<th>Error Stat</th>
<th>CDF percentiles</th>
<th>I-E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Frame width</td>
<td>Patch size</td>
<td>T-D</td>
</tr>
<tr>
<td>1</td>
<td>41</td>
<td>23</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>81</td>
<td>23</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>23</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>81</td>
<td>23</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>41</td>
<td>23</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>81</td>
<td>21</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>81</td>
<td>23</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>41</td>
<td>17</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>81</td>
<td>23</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>81</td>
<td>23</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>
In the previous experiments, we compensate by perturbing the true positions of features by a small distance. A displacement is applied to every point randomly with a certain distance. Then, a single test iteration is performed to locate the true positions. The test is repeated 10 times, and the average error is recorded.

There are several ways to evaluate local feature detection using so-called in-the-wild methods. One possible way is to use a global face detector to locate the region of the image that includes the face. Another way is to use prior information about the face region (e.g., a bounding box) and initialise the local feature detector relevant to the bounding box. This method makes the local search evaluation independent from the global ones and has been used in the recent Facial Feature Detection In-the-wild Challenge[58].

However, evaluating the detection of multiple faces with local feature detection is complicated and somewhat cumbersome. For example, in multi-face images, the regions needs to be assigned to relevant ground-truth points, and in cases of false positives and overlapping boxes, resolving this relevance is important prior to the local feature detection task.

Since we use an off-the-shelf face detector, we follow the second method by evaluating the RFCLM when initialised to the relevant position of the ground-truth bounding boxes. We initialise the model to two reference points assumed to be the relative position of the box at 25% from the top and side edges. We then fit the full-shape model to best match the initial points of the (external corners of the eyes) (see Figure 3.8 for some examples).

![Figure 3.8: Examples of initialised points produced from prior information of face bounding boxes.](image)
3.5.3 Tracking Evaluation

After determining all the configuration parameters for the RFCLM at all stages and levels, we conduct another experiment to understand how RFCLM performs under video tracking settings. This involves determining a threshold value that is used as a discriminative measure for false and true facial regions. Below, we demonstrate how this parameter can be examined to find optimal RFCLM configurations for video tracking.

For efficient tracking, the multi-stage search strategy is applied only when initialising the search, then the RFCLM is used at the last stage to track local features. Previous stages are used when necessary, such as in cases when tracking is lost due to significant changes in facial poses or expressions. Repeating the search with previous stages, or even with the global detector, is only used when necessary to correct subsequent local search regions. Therefore, it is essential to determine whether the current search region of RFCLM is considered to be a true or false positive.

In the RFCLM[22] method, when searching for local features, each local detector produces votes of predictions using trained random forest (RF) decision trees. This creates a cost image as shown in Figure 3.9 in which an optimisation function is applied to find the best match of both the likelihood of the shape and the maximum votes, which produces a quality of fit measure that plays an important role in our tracking experiments as illustrated below.
3.5. **RFCLM OPTIMISATION**

We observe the quality of fit in every RFCLM model when evaluated under different initialisations using the Helen\[56\] dataset (see section 3.5.1 and 3.5.2). Our goal is to seek an RFCLM model with a threshold that maximises the number of passed samples under a certain threshold with the lowest number of prediction errors. To compare different models, we use quality of fit at the 95% percentile of each model, computed from samples with mean error less than 10 relative distance.

We evaluate all models on the Denver Intensity of Spontaneous Facial Action (DISFA)\[2\] dataset, which contains several video sequences of 32 subjects annotated with 66 facial landmarks as shown in Figure 3.10. In our evaluation, we use a 68-point RFCLM, and the error measure is computed based on the 66 points correspondences of multi-PIE\[61\] annotation scheme.

We find that the top results, as shown in Figure 3.11 that were produced using various thresholds computed from evaluations with different initialisation errors, achieve the
same results, indicating the robustness of the measure. A statistical summary of the error measure with relevant threshold values is shown in Table 3.9 and Table 3.8, respectively.
Figure 3.11: Top RFCLM Tracking Results on DISFA dataset when used with a tracking threshold set to the 95% percentile value of the quality-of-fit.
Table 3.8: Distances less than 10% of reference length. I-E refers to Initialised Error Distance either in a relative distance or pixels (px)

<table>
<thead>
<tr>
<th>I-E</th>
<th>Frame width</th>
<th>Patch size</th>
<th>T-D</th>
<th>Scale factor</th>
<th>Quality of fit (percentiles)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>median</td>
</tr>
<tr>
<td>10px</td>
<td>161</td>
<td>17</td>
<td>15</td>
<td>0</td>
<td>166</td>
</tr>
<tr>
<td>10%</td>
<td>161</td>
<td>17</td>
<td>15</td>
<td>0</td>
<td>166</td>
</tr>
<tr>
<td>20%</td>
<td>161</td>
<td>17</td>
<td>15</td>
<td>0</td>
<td>177</td>
</tr>
<tr>
<td>30%</td>
<td>161</td>
<td>17</td>
<td>15</td>
<td>0</td>
<td>185</td>
</tr>
<tr>
<td>10px</td>
<td>161</td>
<td>23</td>
<td>7</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>10%</td>
<td>161</td>
<td>23</td>
<td>7</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>20%</td>
<td>161</td>
<td>23</td>
<td>7</td>
<td>0</td>
<td>137</td>
</tr>
<tr>
<td>30%</td>
<td>161</td>
<td>23</td>
<td>7</td>
<td>0</td>
<td>141</td>
</tr>
<tr>
<td>10px</td>
<td>41</td>
<td>19</td>
<td>15</td>
<td>1</td>
<td>68.4</td>
</tr>
<tr>
<td>10%</td>
<td>41</td>
<td>19</td>
<td>15</td>
<td>1</td>
<td>68.4</td>
</tr>
<tr>
<td>20%</td>
<td>41</td>
<td>19</td>
<td>15</td>
<td>1</td>
<td>68.8</td>
</tr>
<tr>
<td>30%</td>
<td>41</td>
<td>19</td>
<td>15</td>
<td>1</td>
<td>69.5</td>
</tr>
</tbody>
</table>

Table 3.9: Top RFCLM under video tracking settings.

<table>
<thead>
<tr>
<th>RFCLM configurations parameters</th>
<th>Maximum point error (percentiles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Frame width</td>
</tr>
<tr>
<td>----</td>
<td>-------------</td>
</tr>
<tr>
<td>1</td>
<td>161</td>
</tr>
<tr>
<td>2</td>
<td>161</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
</tr>
</tbody>
</table>

3.5.4 Final Model Evaluations

The final RFCLM model is trained on a total of 5,643 images from all datasets mentioned in 3.1. We include the reflection of each sample, thus doubling the number
of training samples to \(= 11,286\) images. We evaluate the optimised RFCLM under a number of public datasets allocated for testing. We compare RFCLM performance on 300W-2 dataset [3] as shown in Figures 3.12, 3.13 and 3.14.

Figure 3.12: 51 points RFCLM comparison with results of 300W challenge [3] on indoor and outdoor images of 300W-2 dataset.
Figure 3.13: 51 points RFCLM comparison with results of 300W challenge [3] on indoor images of 300W-2 dataset.
3.5. RFCLM OPTIMISATION

Figure 3.14: 51 points RFCLM comparison with results of 300W challenge [3] on outdoor images of 300W-2 dataset.

Table 3.10: A comparison with several state-of-the-art methods on 300W dataset, which is divided into two subsets: common subset (containing both HELEN and LFPW test subsets) and challenging subset (IBUG dataset)

<table>
<thead>
<tr>
<th>Method</th>
<th>Comm.</th>
<th>Challenge</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xiao et al. [21]</td>
<td>4.12</td>
<td>8.35</td>
<td>4.94</td>
</tr>
<tr>
<td>Feng et al. [20]</td>
<td>3.27</td>
<td>7.18</td>
<td>4.04</td>
</tr>
<tr>
<td>Lv et al. [63]</td>
<td>4.36</td>
<td>7.56</td>
<td>4.99</td>
</tr>
<tr>
<td>Optimised RFCLM</td>
<td><strong>5.14</strong></td>
<td><strong>14.9</strong></td>
<td><strong>7.05</strong></td>
</tr>
</tbody>
</table>

Our results show that RFCLM achieves comparable results on all datasets. However,
since the initialised boxes of the 300W-2 [58] are not publicly available, we aimed to replicate the initialised error by computing the bounding box of the ground-truth and initialising the model to a fixed relative position to the box. In our test, the resulting initialised error is slightly larger than the one used in the competition. For example, in the 300W[58] challenge, 30% of the test samples are initialised with a mean error rate less than 15% of relative distance, while in our case, the initialised errors are greater than 15%. We compared our results to several recent state-of-the-art methods published in 300W challenge [3]. We also show statistical comparisons on 300W dataset as shown in Table 3.10

3.6 AAM Optimisation

Subject-specific AAMs (i.e., models trained on images of a single individual and then applied to other images of the same individual) are known to be robust for tracking even in difficult and noisy conditions. In this section, we describe how we build an AAM model and optimise some of its parameters. The purpose of building this model is to use it for a set of specific subjects (detailed later in chapter 6) whose video sequences are recorded in challenging conditions.

3.6.1 Feature Extraction

Originally, AAMs worked directly on pixel intensities. However, as shown in [64], AAMs can be made more robust when they work on feature images if they are designed to minimise the effects of lighting changes. In this work, we use three plane images in which the first plane is a local z-normalised version of the original image, augmented with the x and y gradients of the image. The z-normalisation transforms each pixel as follows:

\[
g(i, j) \rightarrow \frac{g(i, j) - \mu(i, j)}{\max(\sigma(i, j), \sigma_i)}
\]  

(3.2)
Table 3.11: Examined AAM configuration parameters

<table>
<thead>
<tr>
<th>Reference frame width</th>
<th>Smooth width, $w_1$</th>
<th>Gradient smooth width, $w_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>7</td>
<td>1.5</td>
</tr>
<tr>
<td>31</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>41</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>91</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where $\mu(i,j)$ is a local estimate of the mean, computed by smoothing the original image with an exponential filter with half-width $w_1$, $\sigma^2(i,j)$ is a local estimate of the variance and $\sigma_t$ is a lower bound chosen to avoid dividing by zero. The variance, $\sigma^2(i,j)$, is computed as

$$\sigma^2(i,j) = \hat{g}^2(i,j) - \mu^2(i,j)$$

where $\hat{g}^2(i,j)$ is an image created by smoothing the squared image ($g(i,j)$) with an exponential with half-width $w_1$.

The gradient information is computed by applying Sobel gradient operators, then using an exponential distance function with half-width $w_2$.

### 3.6.2 Parameters and Scope

When training the AAM model, images are projected in a defined reference frame where a correlation is learned from the perturbation of image texture and model parameters. In our experiment, we aim to identify the optimal setting of several parameters, such as the reference frame width and the smoothing widths $w_1$, $w_2$ of the exponential kernels. We examine a range of parameters as shown in Table 3.11, which yields a list of 200 combinations.

Similar to RFCLM, we build a two-stage AAM-based model with multiple levels. The
Figure 3.15: An example of 27 facial landmarking scheme used for annotating samples of AAM experiment. First stage AAM are used to locate 4 points (in red). initial stage is dedicated to locating 4 salient features as shown in Figure 3.15 (red coloured), which is used to initialise the second stage model with 27 landmarks.

3.6.3 Dataset and Experiment

To train the AAM model, we annotate a set of images for four control subjects with 27 facial landmarks as shown in Figure 3.15. Video frames were extracted randomly, between expression onset and offset, for all facial expressions involved in this study.

To reduce the training time, we optimise the AAM parameters for a single subject and apply the optimal settings for all other subjects. We use a total of 108 images for the benchmark subject (74 images for training and 34 for testing). We also double the sample size by using reflected images.

To train the models, we sample at 50 randomly displaced positions around the points on each training image. We displace by up to 10% in x,y, 10% in orientation/scale, and 0.5 standard deviations for each shape parameter.

To select optimal models, we rank the results based on the median of the CDF function of the mean distance error over the points. The final results of stage 1 and stage 2 are shown in Table 3.12 and Table 3.13, respectively.
Figure 3.16: Top subject-specific AAM models based on median at first stage. Figure keys represent models frame with (e.g., a model with frame width of 101 pixels is denoted by FW101). The model with the minimum median is highlighted with a red circle in the Figure while parameters values are shown on under the title.
Figure 3.17: Top subject-specific AAM models based on median at first stage. Figure keys represent model frame width (e.g., a model with frame width of 101 pixels is denoted by FW101). The model with the minimum median is highlighted with a red circle.
3.6. AAM OPTIMISATION

Figure 3.18: Top subject-specific AAM models based on median at third stage. Figure keys represent model frame width (e.g., a model with frame width of 101 pixels is denoted by FW101). The model with the minimum median is highlighted with a red circle.
### Table 3.12: Top Models for 4-points AAMs

<table>
<thead>
<tr>
<th>Level</th>
<th>FW</th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>Mean</th>
<th>SE</th>
<th>SD</th>
<th>Median</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31</td>
<td>3</td>
<td>1.5</td>
<td>3.49</td>
<td>0.080</td>
<td>2.07</td>
<td>2.6</td>
<td>7.06</td>
<td>7.86</td>
<td>8.93</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>7</td>
<td>2.5</td>
<td>3.49</td>
<td>0.086</td>
<td>2.10</td>
<td>2.56</td>
<td>7.07</td>
<td>7.93</td>
<td>9.28</td>
</tr>
<tr>
<td>3</td>
<td>61</td>
<td>11</td>
<td>2.5</td>
<td>5.16</td>
<td>0.136</td>
<td>3.32</td>
<td>4.20</td>
<td>10.10</td>
<td>12.00</td>
<td>14.00</td>
</tr>
</tbody>
</table>

### Table 3.13: Top Models 27-points AAMs

<table>
<thead>
<tr>
<th>Level</th>
<th>FW</th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>Mean</th>
<th>SE</th>
<th>SD</th>
<th>Median</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>91</td>
<td>3</td>
<td>3</td>
<td>3.63</td>
<td>0.092</td>
<td>2.24</td>
<td>2.78</td>
<td>6.79</td>
<td>8.87</td>
<td>12.0</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>3</td>
<td>1</td>
<td>3.52</td>
<td>0.131</td>
<td>3.21</td>
<td>2.65</td>
<td>5.91</td>
<td>6.64</td>
<td>14.7</td>
</tr>
<tr>
<td>3</td>
<td>91</td>
<td>3</td>
<td>3</td>
<td>3.38</td>
<td>0.111</td>
<td>2.71</td>
<td>2.66</td>
<td>5.43</td>
<td>6.43</td>
<td>11.5</td>
</tr>
</tbody>
</table>
As shown in Table 3.13, for the first level, a frame width of 91 shows the best performance in median. For the next level, the top performance is shown in a model with a frame width of 61, which reduces the median error slightly but increases the error rate to the 99th percentile. For the third level search, the top performance is shown again with the same configuration settings of the first model. The median error rate increased in the final level by 1% compared to the second level. However, the overall error is reduced at the 99% by 0.5% when compared to the first level.

3.7 Discussion and Conclusion

In this chapter, we optimised two types of facial trackers: generic RFCLM and person-specific AAM. The optimal parameters were highlighted for both trackers. We evaluated RFCLM on a number of publicly available datasets and we did not outperform the current state-of-the-art methods. However, our optimised RFCLM achieves comparable results on several recent datasets. Most state-of-the-art methods are CNN-based. However, in contrast to CNN-based approaches, our approach is less complex and more efficient. These optimised trackers will be used in Chapter 6 for the analysis of facial masking effect in PD.
Chapter 4

Facial Behaviour Analysis

4.1 Introduction

In natural human-to-human interactions, affect inferences are derived from a multi-modal process, which involves the interpretation of multiple behavioural signals verbal and non-verbal. These communication signals can be interpreted as the reflection of the human mental state or feelings[65]. Non-verbal behaviour can be described by physical activities different from speech such as facial expressions, body gestures, and leg movements[66].

An important part of human non-verbal communication is facial expression. The human face is regarded a major system for signalling emotions[67]. People use facial expressions to regulate interactions and communicate emotions, display comprehension, and express agreements and disagreements. An analysis of verbal and non-verbal messages of feeling and attitudes during the interactions of a group of people revealed that the majority of these messages (55%) were conveyed through facial expressions while 38% were conveyed by vocal signals (the way words are spoken) and 7% by verbal signals (words being said) [68]. Thus, facial expressions provide significant visual cues to human facial behaviour.
Analysing human behaviour from multiple signals has been applied in many behavioural studies and recent research on affective computing is shifting toward this direction [69, 70]. A possible reason for this shift is that multimodal system has many advantages such as wider application domain, extensibility of local models and ability to customise and adapt to specific applications. A recent example is a Multisense framework introduced by Stratou and Morency [69], which exemplifies the conceptual and technical aspects for developing such systems. This involves identifying application-relevant models, tackling synchronization of multiple data streams and providing a control over frame rate. Analysing features derived from multiple behavioural models has been shown to improve the understanding of human behaviour [69, 70]. For example, in affect recognition, when combining eye gaze and speech, the recognition rate improves by 6.13% for arousal and by 1.62% for valence [70].

AMFE have been attracting attention since the 1970s. According to Bettadapura [71], the first attempt to build an automatic facial expression recognition system was presented by Suwa et al. in 1978. However, research interest in recognising facial expression did not become widespread until the early 1990s [71]. This lack of interest was ascribed to insufficient automatic facial detection and tracking at that time. However, with the great advances that have been made in terms of hardware and software, which enable faster and more robust facial detection and tracking, interest in behavioural studies has started to grow significantly in multidisciplinary areas. In recent years, AMFE has become a multi-domain active research topic. This can be attributed to its potential benefits in various domains, such as affective computing, human-computer interaction (HCI), psychology, neuroscience, and many others. The next section will highlight some of the major potential applications in some domains and demonstrate how automatic modelling of facial expression can be beneficial in these domains.
4.2 Applications

4.2.1 Non-health applications

In 1995, Picard coined the term “Affective computing”, which implies enabling computers to perceive and express emotions [72]. Picard described the importance of emotions and their influence on affective computing. Some researchers argued that synthesising human-like emotions would demand a cognitive appraisal before affecting any arousal. In contrast, many previous studies showed that the arousal of an affect does not always require cognitive appraisal[73]. Picard provided several examples and scenarios to show how computerised human-like features such intelligence, creativity, rational thinking, and decision-making, cannot be achieved without considering human emotions, e.g. creating a virtual effective tutor that can adjust the learning process towards user needs.

Facial expressions are also analysed in empathy research [74] in order to gain a better understanding of emotional reactions. Many studies used a visual stimulus that resembles a human face with various expressions in order to study responses of participants and correlate various measurements with different aspects of empathy such as perception or a certain emotion. Moser et al. [75] showed that the use of auto-generated avatars as visual stimuli allows fine control over facial expressiveness and can elicit emotional responses to similar to human face images, to some degree.

Another study conducted by [76] focused on empathy synthesis projected on an animal robot face. The study aimed to evaluate human emotional responses when exposed to a set of synthesised expressions. The projected emotions were mimicked by an automatic tracking of a set of human facial features mirrored back to the robot face.

Another example can be seen in the safety-critical context, such as in car driving and nuclear plants, where an alert system can be beneficial. An assistive tool can be built to analyse the driver’s facial expression and signal alarms based on signs of fatigue[77]
4.2. APPLICATIONS

or drowsiness[78]. AMFE enables affect synthesis which can be beneficial for domains such as research and entertainment.

In emotion research, visual stimuli are often used to elicit emotional responses. Automatic synthesis of facial expressions can help generate a visual stimulus using avatar images. Since synthesis implies parameterising facial dynamics, these parameters provide control over stimulus intensity. Another example was provided by [79] who developed a simulation tool for generating visual stimulus of facial expressions by using parameters of FACS action units. This allows researchers to control facial muscle movements on every generated example. Affect synthesis is also used in the field of entertainment for modelling virtual faces (avatars) of any kind of character with human-like emotions. These avatars can be used in gaming to improve user interaction experience with more intelligent system interfaces. Animated avatars can also be useful in movie development. They enable efficient movie production with less human action, which can be prone to error.

In marketing for example, automatic facial expression analysis can support marketers by creating more dynamic and effective promotions. The analysis of customer behaviour from facial expression and their responses can provide significant cues to promotion efficacy. Such a study was conducted by [80] where more than 1500 videos of subjects watching a number of commercial advertisements were collected with their responses towards the videos. Their study showed that evaluation of spontaneous smile using facial expression can be a promising indicator of advertisement effectiveness.

4.2.2 Health Applications

In the field of medicine, AMFE can also be used as a descriptive tool for some facial-related medical conditions. The main motivation behind adopting these tools can be ascribed to their ability to recognise or sense certain affects or behaviours. For
example, an assistive tool could be beneficial for monitoring patients with speech disability and recognising certain emotional states such as severe pain, depression or anxiety[81, 82].

In neuroscience, it is well known that the amygdala, a group of brain neuron cells, plays a significant role in human perception and production of emotional facial expressions. However, the role of the amygdala in response to specific emotional facial expressions is not clearly understood and is the subject of many studies.

Previous studies in neuroscience suggest that certain types of emotional responses that are automatically elicited do not require subject awareness. As suggested by [83], when viewing a target stimulus of an angry face for a very short time masked by a neural face (conditional masking), the subject would only report seeing the mask and not the target, while under the conscious level, there are still evidently elicited emotional responses.

Another similar study was conducted by Duan et al. [84], who analysed neural activities of the amygdala using brain scan signals to evaluate subject’s responses when viewing surprised and happy facial expressions. In their experiments, visual stimuli were presented using a conditional backward masking procedure.

If the findings of such studies are proven, they can be helpful in evaluating the function of the amygdala. While using brain scan signals is known to be a robust procedure, it is costly and time consuming. Since there are some known diseases that can affect the cognitive functions and emotional processes of humans, AMFE can be used to assess patients’ facial expressivity and evaluate their efficacy in capturing their deficits by quantifying facial expression impairments.

In clinical studies of facial expression, it has been proven that certain diseases such as Parkinson’s and schizophrenia are associated with impairment of facial expression. Thus, if these impairments could be assessed automatically, this will improve clinical assessment, allow clinicians to monitor daily disease progression, and enable early...
medical intervention.

4.3 Definition of Facial Behaviour

Prior to detecting facial expression deficits, one has to define their normality and how they are developed. Historically, many studies have attempted to define human facial behaviour. These attempts began since the 17th century; in 1649, Bulwer defined several facial emotions and their corresponding facial muscle movements in his book *Pathomyotomia* [71]. Moreover, in the early 19th century, Charles Bell published an extensive structural description of various emotional expressions and their associated respirations (e.g. how muscles around the eyes are contracted involuntarily in violent expressions) [85].

Affect is defined as an expressed emotion or feeling [86]. Facial expressions convey emotional states (affective states)[87], which are manifested by changes in the human facial muscles[88]. This has led many studies to analyse the relationship between these movements and their corresponding affects using various scientific measures [87].

A number of emotional states can be produced by facial muscle movements. Thus, many attempts have been made to conceptualise these emotions [89, 90]. Despite these efforts, researchers have not yet reached an agreement on the conceptualisation of emotions [91]. In the context of facial behaviour analysis, the terms *affect* and *emotion* are not clearly distinguished and are interchangeably used in the literature [86, 92].

4.3.1 Emotion Theories

How humans attained emotions is a highly debated topic [93]. Some claims suggest that emotions are an evolutionary development and learnt from cognitive appraisals [94]. In cognitive theories, the *Component Process Model* (CPM) describes the arousal of an emotion as an event resulted from subjective appraisals [95], where many processes are
automatic (under the conscious level). These processes include cognitive functions such as attention, motivations and evaluations of context surrounding the event. A recent review of cognition and emotion research development can be found in [96]. The review discusses how cognition impacted emotional research and highlights major development of the field, particularly in appraisal theories.

On the contrary, many have rejected these claims and emphasise that emotions are independent of cognition [97]. These concerns have led to questions regarding whether or not emotions are universal. Several theories have been proposed regarding what constitutes affects or emotions [90, 98, 99], which can be classified into three categories: categorical, appraisal-based and multidimensional methods. Appraisal-based methods refer to expressed emotions as results of cognitive processes, which occur prior to the activation of the emotional state. Owing to this fact and since most automatic affect-recognition applications rely on physical effect, appraisal-based methods are considered to be out of the scope of this study. The discussion below will only include both categorical and multidimensional approaches.

### 4.3.2 Categorising Emotions

Concerning categorical methods, one popular approach was presented by Ekman and Friesen [100]; they studied a large number of emotions and presented six basic emotions: surprise, fear, disgust, anger, happiness, and sadness, which are found to be universally recognised regardless of differences in culture. These basic emotions gained a lot of research attention in the automatic affect recognition field [101, 102]. However, others claim that these basic emotions are insufficient for affective computing as they are not used often in the HCI context [103]. It also has been noted that these basic emotions were never intended to represent the range of emotions that can occur in everyday life [87, 104].
Another categorisation of emotions was described by [98], who used a large English-based taxonomy, formed from linguistic analysis of terms used to describe these emotions. This taxonomy contained a complex list of 412 emotional states categorised in 24 groups. However, unlike basic emotions, there was a lack of evidence that these complex emotions are universal [105].

A common standard measure, which is based on morphology and widely used in AMFE, is that of describing facial expressions using the Facial Action Coding System (FACS)[100]. FACS provides a general definition based on the physical morphology of one or more facial muscles in terms of their contractions and relaxations to describe facial expressions.

The use of the FACS scheme involves several drawbacks, which affect its adaptability. Firstly, FACS is prone to errors. The coding of AUs requires expert knowledge and skills [106]. For the early version of FACS, the training time required is approximately 100 hours. Furthermore, a large number of possible behaviours can be encoded by FACS. Thus, an alternative guide of FACS has been proposed [107], which is more subjective and time oriented, with respect to a specific set of emotions. One of the challenges associated with FACS is the estimation of the intensity of Action Units (AUs), which describe the degree of facial expressivity[104].

Another recent work conducted by Jack et al. [108] reduced the six basic emotions to four. A Bayesian classifier was used to discriminate the six basic categories of emotions over a temporal space. An analysis of the relationship between the signals of Action Units (AUs) showed that early confusion occurrences confuse the classifier when discriminating between two pairs of emotions. This can be attributed to the fact that some pairs of emotions are similar, such that they share common signals of AUs. It has been found that anger and disgust both have nose wrinkling effects while happy and surprise share wide open eyes. Thus, they reduced the number of categories by classifying each pair as one category, resulting in four categories. However, further processing was conducted to discriminate between those pairs of emotions.
4.3.3 Multidimensional Space

An alternative approach to describing emotional states is that of representing them as points in a multidimensional space \[99, 109\]. The representation of these dimensions usually accounts for similarities and differences between these emotions \[99\]. One example is a theoretical model called the wheel of emotions \[35\], which describes the relationship between various emotions. The wheel model displays eight primary emotions - fear, surprise, joy, trust, sadness, anticipation, anger, and disgust. Similar and opposing emotions are described in a 2D circle or 3D cone, placed in pairs of polar opposites as shown in Figure 4.1.

![Figure 4.1: Plutchik’s (1980) wheels of emotions[4]](image)

The intensity of each emotion increased when moving to the dark areas and vice versa. As the colour mixture would usually result in similar or different colours, the same applies to this model. Emotions can vary, so a mixture of intensities may result in different emotional states. For example, trust and joy would result in love while trust and fear would result in submission etc.
4.4 Automatic Models of Facial Expression (AMFE)

In this section, we highlight and discuss major challenges in AMFE and provide a brief background to the general steps of AMFE from 2D images. Furthermore, we review some of the recent advances in AMFE with a greater focus on clinical applications for deriving objective measures.

4.4.1 Challenges

There is no doubt that the success of AMFE can lead to a number of opportunities as mentioned earlier in Section 4.2, yet there are many challenges associated with AMFE that need to be addressed. Below, we address the major challenges in AMFE in terms of the dependent processes, context settings, targeted facial expressions, and emotion elicitation procedures.

4.4.1.1 Tracking Challenges

A fully automatic model of facial expression implies the use of a robust automatic facial tracking system in order to locate the face region and associated local features prior to the analysis process. Research in automatic facial tracking systems has improved
significantly since the 1990s; however, the success in this field is still limited in real-world conditions[? ]. This can be attributed to the high variability in the face images in terms of scale, rotation, head-pose, illumination, and occlusion. To understand how these variabilities affect facial tracking, please refer to section 2 for more details. Considering the aforementioned reasons, it is important that prior to the analysis of any expression-related data, one must consider the degree of freedom/constraints and how to account for these variabilities. These variabilities can degrade not only the tracking performance but also the whole process chain of AMFE.

4.4.1.2 Targeted Expressions

The elicitation procedure and the class of emotions are important aspects that can affect the level of difficulty associated with modelling any expression[71]. The elicitation procedure determines whether the elicited expressions are natural (spontaneous) or artificial (posed). Posed expressions are much easier to model than spontaneous ones because they are exaggerated and often frontal (subject facing camera). These usually result in expressions with higher intensity and few head motions. Thus, the learning task of AMFE is limited to a small subset of scenarios when compared to real-world conditions.

On the other hand, spontaneous expressions are hard to model or learn mainly because modelling spontaneous expressions would require a large number of samples and collecting and annotating spontaneous samples is very difficult, time consuming, and usually error prone unless a semi-supervised learning method is used[110].

There are various categorisations of emotions in the literature. One of the most commonly used classifications is that of basic emotions, which refers to several universal emotions identified by Ekman et al. [87], namely happy, sad, anger, fear, disgust, and surprise; for more details, please refer to section 4.3.2. Many studies found that modelling positive expressions such as happy or smiley faces is much simpler than modelling negative expressions like angry and sad faces[71, 111].
The challenges in AMFE can be classified into two groups: facial tracking-based and expressivity-based challenges. Facial tracking-based challenges are driven by high variability in head pose, illumination, and occlusion. Thus, before analysing any expression-related data, it is essential to account for these variabilities.

Expression-based challenges involve noise filtering and extracting expression-related features. The former type can be related to the model’s ability to identify and discard noise data generated from the facial tracking step, which occurs before the analysis of these features. The latter group consists of challenges in discriminating expressive features from other non-expressive features such as those related to illumination, head pose, and inter-subject differences (e.g. identity, gender, age, and ethnicity).

4.4.2 Steps of AMFE

Automatic analysis of facial behaviour usually requires many pre-processing steps in order to collect input features from the face image. The process usually starts with the localising of the face region within the image. After detecting the face region, subsequent processes are applied within the face region in order to extract local features. The extracted features are then used as an input for the interpretation processes. Below, we provide an overview of AMFE steps for analysing 2D images.

The process of AMFE consists of four steps: registering the face image into a reference frame, encoding the expression data, generating expression-sensitive features, modelling, and deriving objective measures. The expression data are usually represented by spatial or spatial-temporal data. In spatial representation, facial expression is encoded using a single frame while in spatial-temporal representation, expression is encoded from a sequence of frames, so the expression dynamics over time such as onset, apex, and offset are included [112].

Approaches in AMFE can be categorised based on their input and output features. The input features that are often seen in AMFE research are either geometry-based,
appearance-based, or a mixture of both\cite{112}. Geometry-based features rely on the geometrical shape of the face using landmarks to describe the geometrical position of facial features while appearance-based features rely on the image texture properties.

The other classification of AMFE techniques is based on the output data: \textit{classification-based} or \textit{regression-based}. The former type is generally used in affect recognition problems where each input is labelled with a certain affect (e.g. a smile) while the latter aims to describe affect using continuous values in a single- or multi-dimensional subspace. The figure 4.2 depicts the general process pipeline of AMFE.

![Figure 4.2: Process pipeline for expression modelling](image)

Face registration is a fundamental step prior to facial behaviour analysis that aims to account for rigid movements caused by head pose and body. The goal of this step is to establish a common reference where samples can be comparable. There are three main categories of face registration: full face, part-based, and point-based \cite{112}. In full face registration, rigid (e.g. affine) or non-rigid (e.g. piece-wise affine) transformation is applied to map the whole face to the reference frame. However, in part-based registration, the mapping is based on sub-images of facial parts (possibly of different sizes) such as eyes, mouth, etc.
4.4. AUTOMATIC MODELS OF FACIAL EXPRESSION (AMFE)

4.4.3 Feature Extraction

To obtain a richer representation of the facial appearance we extract texture information from patches around each model point. We explore different types of texture representation, summarised below.

4.4.3.1 Gabor

Gabor filters are linear filters with several characteristics such as spatial location, orientation and frequency. The main function of these filters is to detect whether there is a signal in a certain direction, orientation and scale. A filter bank containing Gabor filters is widely used in many applications such as image retrieval[113] and computer vision studies, especially for texture classification problems [114–117].

The spatial Gabor filter consists of two parts, a sinusoidal function and a Gaussian envelope which can be computed using:

\[
G(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\exp\left(i(2\pi x + \psi)\right)
\]  

(4.1)

where the first part of the equation is the Gaussian envelope and the second part represents the sinusoidal function. These two functions are defined by a set of parameters such as the wavelength \(\lambda\) of the sinusoidal function, the orientation \(\theta\), the standard deviation \(\sigma\) of the Gaussian envelope and the spatial aspect ratio by \(\psi\).
Table 4.1: A set of pair Gabor kernels with sinusoidal functions. Each pair of images below represent Gabor function with a sin and cosine function respectively.

<table>
<thead>
<tr>
<th>σ</th>
<th>θ = π</th>
<th>θ = 0.25π</th>
<th>θ = 0.5π</th>
<th>θ = 0.75π</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image0.25pi16.png" alt="Image" /></td>
<td><img src="image0.5pi16.png" alt="Image" /></td>
<td><img src="image0.75pi16.png" alt="Image" /></td>
</tr>
<tr>
<td>8</td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image0.25pi8.png" alt="Image" /></td>
<td><img src="image0.5pi8.png" alt="Image" /></td>
<td><img src="image0.75pi8.png" alt="Image" /></td>
</tr>
<tr>
<td>4</td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image0.25pi4.png" alt="Image" /></td>
<td><img src="image0.5pi4.png" alt="Image" /></td>
<td><img src="image0.75pi4.png" alt="Image" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image0.25pi2.png" alt="Image" /></td>
<td><img src="image0.5pi2.png" alt="Image" /></td>
<td><img src="image0.75pi2.png" alt="Image" /></td>
</tr>
</tbody>
</table>

In our framework we aim to examine the effect of expression manifold learning from Gabor features. In particular we extract the magnitude of the central pixel of the Gabor kernel when evaluated at each facial feature position in the reference frame. We repeat this process using various kernels from the Gabor filter bank shown in Table 4.1 and concatenated all resulting features to represent the feature vector. The kernel size is set to match the sampled patch around each point.

### 4.4.3.2 Local Binary Patterns (LBP)

LBP is a non-parametric method for learning local texture structure which has been widely used in pattern recognition applications. It was first introduced by Ojala et al. [118] for texture classification and since then, LBP has been extensively explored and adopted in multiple applications including emotional intensity estimation, such as pain[119, 120] and smile[121]. LBP is also being used in chronic disease diagnosis
such as dementia[122, 123]. LBP-based features are still considered one of the most successful features for analysing the structures of facial expressions[124], mainly because LBP descriptors are invariant against monotone illuminations, simple, and efficient to implement while holding a descriptive power about local structure.

The original LBP operates on a set of pixels $P$, using the (3x3) kernel, which uses a binary thresholding function that compares the neighbourhood pixels around the central pixel to the central pixel (see equation 4.2). This process yields a binary string for each pixel as shown in Fig 4.3, which is then converted into a real number.

\[
LBP_P = \sum_{i=1}^{P} S(g_i - g_c)2^i \quad \text{where} \quad S(x) = \begin{cases} 
0, & \text{if } x < 0 \\
1, & \text{Otherwise} 
\end{cases}
\]  

(4.2)

Where $g_i$ represents the pixel intensity value at the $i^{th}$ neighbour and $g_c$ is the pixel intensity value at the central pixel.

Another extension of this operator was introduced by Ojala et al. [125] involving the use of a circular neighbourhood, enabling multi-resolution sampling by varying the radius $R$ and number of neighbourhood pixels $P$, which made this type of descriptor invariant to scale and rotations (see Fig 4.4).
Furthermore, Ojala et al. [125] combined classified patterns of binary codes with uniform and non-uniform patterns. A pattern is considered binary if it contains at most two transitions from 0 to 1 or vice versa. For example, 00000000, 00000001 and 10000001 contain 0, 1, and 2 transitions, respectively, so they are uniform. However, 11001100 contains 4 transitions; hence, it is not a uniform pattern. In their work, they observed the occurrence of these patterns using various numbers of neighbourhood $P$. Their findings suggest that uniform patterns of the (8,1) neighbourhood account for almost 90% of all patterns. Similar work was carried out by [126] to estimate the rate of uniform pattern occurrence in the FERET database of facial images [127]. Uniform patterns accounted for 90.6% of all the (8,1) neighbourhood patterns and 85.2% of all the (8,2) neighbourhood patterns.

### 4.5 Dataset Structure

#### 4.5.1 Considerations

As mentioned in [71], in general, there is a lack of sufficient datasets of spontaneous facial expressions. However, for the purpose of this study and to the best of our knowledge, there are no publically available datasets for facial expression disorders, regardless of sufficient labels, annotations, emotions, etc. Despite the challenges associated in obtaining such data, further constraints and implications concerned with
patients’ privacy and data protection[128] can be one of the main reasons which prevents the sharing of such data. Therefore, the majority of researchers[129] conducted case studies involving groups of subjects (controls and diseased) with a relatively small sample size. These case studies help determine possible correlations between certain factors of facial expressivity with clinical measures.

In this section, we describe the structure of the dataset used in our analysis of facial behaviour. Since the aim of this work was to analyse the dynamics of facial behaviour with respect to facial disorders, we focused on analysing expression variations with time using video recordings of subjects performing several facial expressions as well as of controls and diseased groups.

Symptoms of neurological disorders such as Parkinson’s disease can negatively impact patients’ daily activities and their overall quality-of-life [130, 131]. Current methods of
clinical evaluation of these symptoms include manual methods (of interview and observation), which are usually conducted infrequently every few months, but the symptoms fluctuate on a daily basis [131]. Daily monitoring may lead to a deeper understanding of disease progression and complexity and enable early medical intervention. Currently, clinical assessments are impractical for the purpose of daily monitoring owing to the limited availability of health-care resources.

*Remote Patient Monitoring (RPM)* or *telemonitoring* implies the use of technology to monitor patients in non-health care settings, which are naturalistic (non-controlled). Monitoring patients at their own home is considered one of the most promising applications of *telemonitoring*[132]. Telemonitoring can enable automatic daily assessment of chronic disease progression by capturing daily fluctuations of objectified measures[133]. Moreover, measures recorded in patient-relevant settings are considered more reflective of the patient’s quality of life than observations in health-care settings[134].

### 4.5.2 Collection Method

In order to keep our evaluations relevant to patients’ real daily life experience and suitable for automatic daily assessment, all recordings were undertaken in participants’ natural settings such as their home or office. All participants involved in this study were given software installed in a PC or laptop of their preference. In every session, the software prompted the participant to perform several facial expressions as in Fig 4.5, while the subject face was recorded by a webcam. They were encouraged to ensure that their face was in the middle of the recording area; otherwise, there were no constraints imposed. Each video represents a recording for a single session of a subject. In each session, there were two types of expression stimuli: visual and textual (label). Three sequences of facial expressions were projected on the screen and the sample subjects were asked to either produce the displayed textual emotion label or mimic the projected human face. In the initial sequence, the subjects were given a set of written emotion labels followed by a similar sequence of images of human faces and lastly, a sequence
of non-emotional expressions (written action labels) (see Figure 4.5).

Each recording was then reviewed and annotated manually with respect to their expression label and their approximate time interval. All intervals started and ended with neutral expressions, which were considered the baseline for our evaluations.

4.5.3 Parameter Definitions

Before we describe the geometrical shape analysis, we will define all associated parameters.

Let an $n$ number of subjects be denoted by $s = \{s_0, ..., s_i, ..., s_{n-1}\}$, where $i$ represents the index of the $i^{th}$ subject. Similarly, let $k$ be the number of unique expressions denoted by $e = \{e_0, ..., e_j, ..., e_k\}$, where $j$ represents the index of the $j^{th}$ expression and $e_0 = \text{neutral}$. We also denoted the derived feature vector of a single frame by $f$ and all features are denoted by the matrix $F = \{F(s_0, e_0), ..., F(s_i, e_j), ..., F(s_{n-1}, e_k)\}$ where $F_{s_i, e_j}$ represents all features observed from expression $e_j$ of subject $s_i$. In each $F_{s_i, e_j}$ there were $a$ number of observations such that $F(s_i, e_j) = \{F(s_i, e_j, 0), ..., F(s_i, e_j, q), ..., F(s_i, e_j, a)\}$ where $F(s_i, e_j) \subset F$. For every $q^{th}$ observation, there is a $y$ number of samples such that $F_{s_i, e_j, q} = \{f(s_i, e_j, q, 0), ..., f(s_i, e_j, q, y)\}$.

Below, we describe our method to analyse facial behaviour using various geometrical and appearance features.

4.6 Analysis using Geometrical Features

4.6.1 Modelling Intra-subject variation

As for every feature $f$, there was a corresponding vector $x$, which described the 2D geometrical position of facial landmarks (described in section 2.1). To model legitimate shape variations and reduce the vector dimensionality, all tracked facial landmarks were transformed into a relevant statistical shape model (SSM), by using the inverse of (4.3)
\[ x = T(\bar{x} + \Phi b; t) \] (4.3)

This maps all shapes onto a reference frame by \( T(\cdot; t) \), where the transformation parameters \( t \) are scale, orientation, and translation. Now that all non-uniform shape variations can be described using the parameter \( b \), for geometry-based analysis, we used the parameter \( b \) as the feature vector \( f \), so for simplicity, we replaced \( f \) with \( b \) and the matrix \( F \) with \( B \).

Since, we aimed to analyse the dynamics of facial muscle contractions and relaxation, we excluded dimensions of rigid movements such as that caused by different head-pose from the vector \( t \) by fixing their values at the mean which are essentially zeros. To account for inter-subject shape differences, we subtracted the identity shape \( d \) from the shape vector \( b \), which represents the mean of the neutral expression for every subject. However, since the value of \( d \) plays a significant role in determining subject baseline, we took several steps in order to verify its validity. Identity was only derived from features of neutral expression \( B(s_i,e_0) \); we assumed that if most tracked features were reasonably accurate, outliers were expected to deviate from the majority of the data. Therefore, to account for those outliers we applied sorting and one-tail trimming based on normalised Mahalanobis distances.

It is well known that the geometrical features are sensitive to noise. In order to account for such noise, all elements of \( B \) were smoothed (in time) using an adaptive Gaussian kernel to account for varying density caused by different video frame rates. The width of the Gaussian kernel was set to cover a third of a second.

Given the \( n \) shape parameter vectors from the neutral expression sequence for an individual, \( \{b_i\} \), we estimate the vector for the neutral identity using a robust trimmed estimate of the mean as follows:
1. Compute the mean shape $d = \frac{1}{n} \sum_i b$
2. Compute the distance of each shape from the mean, $d_i = |b - d|$
3. Sort the vectors by $d_i$ and discard the 30% furthest from the mean
4. Compute a new estimate of the mean from this subset

Then, all elements of $\delta B$ could be computed as the residuals of $B$ from the identity shape as in (4.4).

$$\delta b = b - d$$ \hspace{1cm} (4.4)

### 4.6.2 Modelling the Expression Subspace

The vector $\delta b$ now describes the displacement from the identity shape in the shape model parameter space. In order to examine each expression we construct an expression subspace for each expression and each individual. Given a set of shape offset vectors making up the video sequence for a particular expression, $\{\delta \vec{b}\}$ we first compute the trimmed mean (discarding 10% outliers), $\bar{\delta \vec{b}}$, then compute the major eigenvectors of the covariance of the vectors about this mean, $Q_e$. This allowed us to describe the variation using parameter $c$ as shown in (4.5).

$$c = Q_e^T (\delta b - \bar{\delta b})$$ \hspace{1cm} (4.5)

Where $Q_e$ is a set of eigenvectors describing directions of major variations with respect to expression $e$ and $\delta b$, representing the computed mean from all vectors of $\delta b$ in $\delta B_e$. To increase the robustness when computing $\bar{\delta b}$, we apply the same sorting and trimming as described earlier; however, the trim size was set to 10% of the original data.
4.6.3 Neutral Expression Distribution

The activation of any expression state is usually considered at some point during the subject’s transition between current and target expression. Therefore, in order to measure certain expression intensity, it is important to understand their distribution with respect to relevant expressions. This enables further statistical measures to be established (e.g. mean, confidence of estimate, etc.) either locally (per subject) or globally (for all subjects).

In this study, all subjects’ recordings begin and end with a neutral expression. Thus, we aim to estimate the distribution of neutral expression by projecting all vectors of $\delta b \in \delta B_{e_0}$ using equation (4.5) with every expression subspace defined by $Q_{e_j}$ where $e_j \in e$ and $j \neq 0$. This will result in a set of neutral samples denoted by $V_{e_j}$. Since the identity shapes resemble the neutral expressions which are subtracted in (4.4), all resulting parameters of $v$ are expected to be around the origin. Furthermore, in order to examine the nature of the distribution (whether it is Gaussian or not), we modelled all the healthy group data (group details will be discussed later in the case study chapter) in order to visualise the distribution of the data as shown in Figure 4.6.

It can be seen in Figure 4.6 that the distribution of the neutral expression in several subspaces of emotions is definitely not a Gaussian distribution, confirmed by a Mardia normality statistical test as shown in Table 4.2, meaning that all subjects do not share a common baseline in those subspaces. This is probably due to the biased samples, illustrated in the next section.
4.6. ANALYSIS USING GEOMETRICAL FEATURES

(a) anger 27-CLM

(b) happy 27-CLM

Figure 4.6: Neutral data projected in several expression subspaces
Table 4.2: Mardia normality test

<table>
<thead>
<tr>
<th>Expression</th>
<th>Test</th>
<th>Statistic</th>
<th>p-value</th>
<th>Normal?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>Skewness</td>
<td>24025</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>Anger</td>
<td>Kurtosis</td>
<td>284</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>Happy</td>
<td>Skewness</td>
<td>25529</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>Happy</td>
<td>Kurtosis</td>
<td>265</td>
<td>0</td>
<td>NO</td>
</tr>
</tbody>
</table>

4.6.4 Reducing Effect of Bias

Reducing dimensionality using PCA is sensitive to outlier samples which can affect the orientation and the unit distance of the principal axes[135]. Since an approximate time interval is used to extract relevant expression frames and owing to the fact that all videos have varying frame rates, built PCA models based on such samples are likely to be vulnerable to the subject-biased or neutral-biased problem. The term subject-biased means the weight of the distribution of all samples across subjects is significantly higher for certain subjects than others, which would result in a model less effective at representing the average variation across the whole group. On the other hand, by neutral-biased, we mean that the number of frames with neutral expression is significantly higher than the target expression, leading to a model describing the major variation in the neutral expression rather than the target expression. For instance, in a section of video which is supposed to contain the subject demonstrating a "happy" expression, only a minority of the frames actually include the subject smiling - most of the frames are neutral expressions at the start and end of the sequence. To account for the aforementioned problems, we carried out the following steps:

- Reducing the size of neutral samples prior to modelling the expression’s subspace.
- Excluding extremely short or incomplete expressions structure.
- Excluding samples that do not reflect the target expression.
- Refining the annotated expression’s time interval in every video to reduce the
In order to distinguish between neutral samples and non-neutral ones, it is essential to estimate their relative distributions. Since the distribution of neutral space is non-normal, we used a mixture of Gaussian probability density functions (PDFs) to model every subject’s neutral subspace \( V(s,e_j) \). Then, we estimated \( p(c \in V(s,e_j)) \) and excluded all expression samples that fell within a threshold \( t \) (e.g., \( t=98\% \)) of the total distribution of neutral samples, which can be computed using equation (4.6).

\[
Z(c) = \begin{cases} 
  \text{true,} & \text{if } \sum_1^K w_k p_k(c \in V(s,e_j)) \leq t \\
  \text{false,} & \text{Otherwise}
\end{cases}
\] (4.6)

Where \( K \) is the total number of components in the Gaussian mixture for a single subject and \( w_k \) is the associated weight of the \( k^{th} \) component and that \( \sum_1^K w_k = 1 \).

Since our aim is to analyse a complete structure of facial expression dynamics, we discarded samples with short tracking duration less than 1.5 seconds (more than 50% of tracking loss at least). In the next step, we aim to define the the expression time interval from the onset till the offset. A sample of data is shown in Figure 4.7 with approximate targets for onset and offset.

### 4.6.5 Onset and Offset Detection

In order to detect the onset and the offset, we examined the change of the first model parameters which describe the most variation of the data. We examined the gradient of the time curve by computing \( g(t) = \delta b(t) - \delta b(t - \delta t) \), where \( \delta t \) is in the range \([0.1,0.5]\) secs. Peaks and troughs in this curve correspond to the onset/offset of the expression. We noticed that a time step of 0.5 seconds gives a good initial estimate of onset and offset locations as shown in Figure 4.8. Then we refine the position by following gradient.
Figure 4.7: Sample of expression model parameter variation overtime. Shaded areas show optimal target points for onset and offset.
change opposite to the expressivity direction (positive/negative) for the onset while for the offset we follow the gradient in the expressivity direction. The Figure 4.8 shows original curve with various gradient curves using different time steps $\delta t$.

![Figure 4.8: Detection of expression onset and offset using various frequencies kernels](image)

4.6.6 Refining Data

CLM is known to be robust in the local search for facial features in frontal images[22]. However, the success of CLM is dependent on a good initialisation and in our case, since we track subject faces in naturalistic settings, cases with false samples when RFCLM is initialised on a false region that does not contain a real face are highly possible. This can lead to false positives input fed to the expression modelling process. In order to eliminate falsely tracked points from the expression modelling process, we analysed the tracking behaviour in a series of consecutive frames and compared the tracking
behaviour in it. This was done in order to identify possible discriminative features. We observed the variations of the parameters $\Delta \mathbf{c}$

Inconsistent Shapes

When the tracker is initialised on a region which does not contain a face, the first model parameter varies rapidly between frames and shape was not stable. This high variation was used as a key indicator for noisy frames. We classify noisy frames based on a threshold applied on the change of the first element of $\mathbf{c}$ in time when $\delta t = 0.1$. We computed the total number of classified noisy frames in each expression and discarded expression data with a noise larger than 10% of the original data.

4.6.7 Normalising Samples

Expression samples can vary in length, either in time, frame rate or the number of tracked frames. Therefore, to model the shape of the expression curve we apply a normalisation on the size of collected frames in each expression interval to contain only 50 frames as shown in Figure 4.9. This results in comparable expression samples, which can be analysed further using PCA method to model the variation of the expression shape.
Figure 4.9: Frames-normalised expression data. The Figure shows samples of two subjects from two groups (diseased and control) performing several happy expressions. The diseased group data are red coloured.

4.7 Conclusion and Discussion

Our aim in this thesis is to produce a model with parameters that can describe an objective measure based on facial behaviour that will correlate with facial expression impairments. The success of AMFE is dependent on tracking performance and the quality of the sampled data. Therefore, in our evaluations consider the following:

- Evaluating the tracking performance on video sequences
- Quantify discarded frames due to noise or incomplete expression
- Record several measures derived from several features.
• Conduct a regression analysis to determine the best parameters that correlate with the disease effect.

Impairment of facial expressivity is found to be associated with many disorders. A recent survey showed that 39 disorders affecting the brain are associated with reduced facial expressivity[129] For this purpose, we examine several geometrical-based and textural-based features as described above and their fusion to estimate several measures of facial expressivity such as the apex(maximum), baseline, the length of the expression, distance from baseline to the apex, the variance and magnitude. Then we apply a regression analysis to examine the correlation of the derived features with patient QoL scores. A summary of this process can be seen in Figure 4.10.
4.7. CONCLUSION AND DISCUSSION

Figure 4.10: A summary of facial behaviour analysis process
Chapter 5

Gaze Estimation

5.1 Introduction

Gaze direction provides very sensitive cues to human behaviour. Studies on Eye Gaze Tracking (EGT) have a long history that can be traced back to the early 19th century [136]. Tracking gaze directions is a multidisciplinary research interest due to its wide application domain.

In behavioural studies, EGT can be used to analyse or assess patterns of eye/gaze movements associated with normal and abnormal behaviour such as emotion recognition, attention, memory, etc. Some cognitive impairments are often observed in multiple neurological disorders such as PD [137, 138], Alzheimer Disease (AD) [139], Autism Spectrum Disorders (ASD) [140], Prosopagnosia (face recognition impairment) [141].

Automatic EGT can also be used as an interface for many applications, by mapping gaze directions to inputs. For example, gaze directions could be used as assistive tools for disabled people to enable data entry. In the gaming industry, EGT systems can be used to improve Human-Computer Interaction (HCI) experience, for instance, a company named SteelSeries introduced a peripheral device named Sentry eye tracker
in 2014, which enables gamers to use their eye movements and fixations as additional control.

Many computer vision applications utilise gaze information to improve system overall performance. For example, Yun et al. [142] attempted to identify and classify image-based salient objects according to human interpretations of these images, using an eye-gaze tracker. In another study, Fathi et al. [143] used human gaze (fixations and saccades) to improve recognitions of daily actions. There is another line of research that is devoted to predicting driver’s visual attention [144–146], enabling driving’s safety-related applications to incorporate such critical information.

Automatic EGT can be beneficial to a wide range of applications [147] and a robust and accurate automatic EGT system can be an essential part for the success of many behavioural-related applications. Below we review general techniques used in EGT. This includes a range of intrusive and remote categories.

5.2 Eye tracking Methods

Methods of EGT can be classified into two categories: intrusive/invasive and remote.

Intrusive methods usually involve attaching sensing devices to subjects, such as skin-electric sensors, lenses, glasses, head-wearable items, etc. Methods under this category are highly robust and accurate, with an error range of 0.5 degrees [148]. However, because of these attachments, they are considered less convenient for many users and not well suited for practical daily monitoring.

These methods can be classified into: Electro-Oculography, Sclera Search Coil, Infrared Oculography and Video Oculography [149]. The term oculography is a general term used to refer to the techniques for tracking the eye movements.

In Electro-Oculography techniques, a sensor is usually attached to the skin around the
subject’s eyes to record electric pulses, which enables eye movement detection; this method can detect eye movement even when the eyes are closed. In Sclera Search Coil method, a small wire coil is attached to a contact lens and allows the eye to move within a magnetic field. If the coil moves, a magnetic field signal is generated, which enables the system to detect and track eye movements.

Methods which utilise Infrared Oculography are called (active methods), as the eyes are illuminated by infrared light, thus, enabling EGT by measuring the intensity of reflected light on the sclera. In infrared-effected images, the pupil and the iris are found to be more distinguishable. Large variations caused by texture and colour can be eliminated with this effect to a certain degree.

The Video Oculography approach implies video recording of the eye image. However, with some variations. For example, this method can be remote (with no attachments to the user) or intrusive (e.g., using a head-mounted camera). It can also be invasive (using active light) or passive. Moreover, systems under this category can have a simple or complex setup, comprising single or multiple cameras or light sources.

On the other hand, remote methods utilise computer vision-based techniques and do not require any kind of attachment to the subject’s body. This category has received a lot of attention in recent years owing to its potential benefits and wider domain of applications [150, 151].

5.2.1 Feature extraction

Computer-based EGT techniques that rely on visual information can also be categorised into two groups: feature-based and appearance-based [152]. Feature-based methods rely on informative local features of the eyes such as pupil, cornea, light reflections etc. Feature-based techniques can also be divided into two categories: model-based and interpolation-based methods [149].

Model-based methods require a global geometric model camera, light source, monitor,
and their geometric information. In this approach the 3D gaze ray is estimated by constructing the optical and visual axes [149]. The optical axes can be represented by a virtual 3D line which intersects with the cornea’s centre and the retina, while visual axes intersects with the pupil’s centre and the fovea as shown in figure 5.1. Gaze is then estimated by the intersection of visual axes with the object in the image. Moreover, in 3D models, the vector pointing from the eyeball centre to the iris centre is considered the gaze direction [153]. An common feature used in model-based approaches is the pupil-glint vector [152], which is the vector from the light reflection point on the cornea to the pupil’s centre.

Interpolation-based methods use a mapping function between facial image features to gaze coordinates, which could be parametric or non-parametric. Calibration data are used to solve unknown coefficients by a numerical fitting process such as multiple linear regression [149]. In appearance-based methods, gaze is estimated based on image content directly using photo-metric appearance[149].
5.3 Challenges

Several challenges have been reported in the literature regarding automatic EGT. Exact and accurate PoG estimation may not be achieved only from the eye image [154]. When viewing an image, the portion viewed in the fovea region (see Figure 5.1) contains a detailed image and PoG can be redirected within about 1 degree inside this region without any eye movement [154].

The 2D image of the eyes alone is insufficient to estimate the gaze direction. An interesting example can be seen in Figure 5.2 which shows illusional faces where the eye positions in both faces are identical. The gaze direction in the left face appears to go toward the right direction, while the right face seems to target its gaze towards the observer. This suggests that gaze direction is not only dependent on the 2D image of the eyes but also on other informative features about head orientation. This has led many researchers to develop algorithms that incorporate head orientation information with features of the eye image.

![Figure 5.2: Taken from Wollaston (1824)](image)

Human eyes can have different sizes, textures, and colours and moreover, the intensities between the pupil and the iris can be very similar and difficult to distinguish. These variations make automatic EGT very challenging. Besides these variations and similarities, occlusions by eyelid coverage, glasses, and scene objects can pose further obstacles to these algorithms.
5.4 Previous Work

There is an extensive body of literature on EGT with various research aims. For example, in [142], gaze information was used along with various object categories such as chair, person, horse, and so on to determine whether or not certain gaze patterns can exist when people look at a specific category of object. An example of this work is presented in [142]. Their aim is to understand the semantics behind object categories by learning how human gaze behaves when viewing images containing these objects. They showed that using gaze fixation information can enhance the object-detection algorithm by reducing the scope of the search to what is important to the viewer, allowing algorithms to understand individual interest in the image [142].

Below we explore only remote EGT, since it has a wider potential than intrusive methods. It is also user-convenient in that it does not require any attachments to be placed on the user. Below, we address several techniques used for tracking eye-gaze direction and their pros and cons.

Kyung-Nam Kim and Ramakrishna [155] presented a model-based method, which predicts the iris centre based on local searching for the longest horizontal distance between the iris edges. Since model-based methods can be very sensitive to poor edges, they used a model of circular shape computed to aid the process of calculating the centre of the iris by fitting the model to the obvious edges [155].

This technique copes with minimal eyelid coverage that affects the upper and lower boundaries of the limbus, and is an invasive approach since a reference is attached to the glasses for detecting head movement. Even though the equations proposed would work only for slight head movements, calculation of additional parameters is required, such as the distance between the eyeball and the projection screen.

Williams et al. [156] introduced an eye-gaze tracker using a Bayesian-based sparse regression model, trained on partially labelled data. Labelled and unlabelled data are acquired during the calibration process. A moving cursor is displayed on the screen,
which stops at certain points, where the user is requested to fixate their gaze on these points and labelled data are generated by capturing the target image and labelled with screen coordinates. The unlabelled data are generated implicitly, while the user follows the cursor motion. In this calibration, 16 pairs of images and screen coordinates are acquired for the labelled set with another 5 unlabelled images between each interval. The input feature vectors in the learning phase constitute a concatenated vector of the image equalised histogram and edge energy information. Temporal indices are also added to the input vector of the labelled set. The missing labels are recovered using sparse coding but incorporate the Gaussian process to estimate the uncertainty of the predictions since the distribution of the missing data would have smooth variation.

The method was tested on 200 images captured at random gaze locations around the computer screen. The reported angular error was around 1.29 degrees. Despite this high detection rate, the field-of-view was very limited to the PC monitor, which would consequently restrict head movements. However, to fully recognise the tracking performance, it is essential to understand the distribution of the test points – how much they vary across the field-of-view.

Sugano et al. [152] used an appearance-based model to estimate gaze directions in an incremental learning framework. The model incorporated head pose information, which used a 3D rigid facial mesh as implemented in [157]. A single eye image was processed with several filters before being used as an input feature. The eye image is wrapped to a reference frame using Delaunay triangulation. Edges of the eye image were highlighted by a thresholding and Sobel filter scan. A smaller patch centred at the edges’ mean point was extracted, which contained the eye image. Histogram equalization and bilateral filters were then applied to reduce the effects of lightening and image noise, respectively. The patch size was further reduced to minimize the reconstruction error of the PCA subspace. The appearance of the eye and the head pose parameters were then vectorised and concatenated into one feature vector, fed
to an incremental PCA to update the subspace to best describe the samples’ PCA coefficients. The used PCA is then clustered based on head pose similarity and each cluster has its own manifold. When testing a new input feature, it gets mapped to relevant head pose clusters determined by cluster-distance measure. Then, gaze is computed as the sum of cluster predictions penalised with a cluster-distance measure for each prediction.

Like most methods, this approach uses only a single eye image as an input feature but the question arising here is which eye image (left or right) should be used and how. The algorithm involves complex image pre-processing and it scores a high average error rate between 4 and 5 degrees. Despite this, the field of view is limited to screen coordinates. It is true that the method avoid initial calibrations, however, implicit calibrations are still performed in the background from the user’s mouse clicks.

Since head movement is a challenging problem for most remote gaze-trackers that do not use an external light source, Sesma et al. [158] conducted a feasibility study on passive EGT using a single camera where no additional light sources were used. As the infrared glint reflection in active methods provides vital cues to head orientation, the study aimed to measure the potential of using eye corner information as a replacement to the glint-reflection. The experiment was performed using only facial features to predict the eye gaze using a single webcam attached under the screen monitor. The gaze is estimated based on a feature vector describing the normalised distance of the pupil centre to the eye corners. A calibration method is applied, using a 4 x 4 screen grid for each eye to incorporate extra calibration coefficient information to the feature vector. Then, the average of the normalised distance between pupil centre and each of the inner and outer corners is used for each eye to finally find the average result for a combined vector for both right and left eyes [158].

The results of the above study showed that the eye corners’ positions are not static when the subject moves their gaze position even when the head pose is fixed. The error rate reported is around 3 degrees with head-pose fixed and a narrow field-of-view
to the computer screen, suggesting that it can increase dramatically under real-world free-head movement scenario. The method was also performed using a high-resolution camera of 1600 x 1200 with 5 fps, which implies that additional high performance cost in terms of hardware and processing time may increase when integrating other tracking applications, which require scaling the frame rate up to more than 5 fps.

Sugano et al. [159] presented a single regression forest trained on both real and synthesised data to predict gaze direction in 2D polar coordinates with regard to camera coordinates. Samples are collected from 50 subjects with several fixed head poses using a chin rest. The setup involved multiple cameras as 160 images were captured from 8 synchronised cameras. Six facial landmarks were manually annotated from the first image of each camera to obtain their 3D locations. These landmarks form the corners of the eyes and the mouth. Their 3D coordinate estimation aids the process reconstruction of the 3D eye image and head pose alignment with the camera coordinates using three midpoints of these corners. The regression forest used head pose-based clusters, trained on concatenated, vectorised eye appearance features and the 3D mid points, which account for the head pose. Input samples are clustered based on their head pose inside the regression forest used the k-nearest neighbour approach. The evaluation of this method was done in two ways: within-subject and cross-subject training and testing. Synthesised data of 33 subjects were tested and showed an average error of around ±4 degrees for within-subject evaluation and ±6.5 degrees for cross-subject.

The above model showed the best performance with synthesised low-resolution images of 15x9 across the dataset being used. However, the complexity of the setup requirements such as the number of cameras and calibration process, prevents its adaptability to more general applications/domains.

There are several approaches that have used Convolutional Neural Network (CNN) methods to estimate gaze directions. Below we highlight the start-of-the-art approaches in this context.
Zhang et al. [160] used an appearance-based multimodal CNN to predict gaze direction in 2D polar coordinates. They aimed to collect 2D samples from arbitrary settings with varying conditions from personal laptops of 15 participants. The dataset contained 213,659 samples annotated with 12 facial landmarks representing the largest dataset in terms of the number of samples with varying conditions. They implemented a background-running application in each laptop using the calibration method of [161] to obtain camera pose parameters and represent the gaze targets relative to the physical 3D coordinates. The facial features were detected using constrained local model [105], while the head pose is estimated by fitting a 3D shape model, initially using the method of [162], then optimised using a non-linear method. The method followed the idea of [159] as described earlier in using head pose-based clusters and camera alignment by parallelizing the head and the camera coordinates. The method was evaluated across several datasets namely Eyediap [163], UT Multiview [159] and MPIIGaze [160] and compared with several state-of-the-art methods. The reported average angular error was 10.5 on Eyediap, 5.9 on UT Multiview, and 13.9 on MPIIGaze.

This work was recently introduced as GazeNet [164] with several improvements in terms of annotations (37,667 samples), network structure, specific-conditions testing and across different datasets. The average angular error was reduced from 13.9 to 10.8 degrees.

Since the CNN method above aims to overcome eye appearance challenges in the normal laptop environment, the distribution of the gaze angles and head poses was limited to targeting a computer’s screen. In cross-dataset evaluations, test samples were selected to represent limited head poses and gaze direction suited to this environment. The method outperforms previous existing methods representing the state-of-the-art in this context.

Another CNN-based model was presented by Park et al. [150]. The CNN is trained on synthetic images of the eye, utilising two geometrical models for the pupil and the iris using circular shapes. Gaze directions is computed as the 3D vector intersecting centres
of the eyeball and the pupil. Eye-corners distances as in [158] have been examined and proved to improve model predictions significantly.

5.5 Summary

Most of the proposed gaze estimators in the literature rely on preliminary input from the user to adopt user-specific parameters, which is usually done during the calibration stage. However, in recent years, passive subject-independent appearance-based models have attracted many researchers in the computer vision community [159, 160] owing to the recent advances in technology that largely expanded the use of camera-attached mobile devices such as smart phones, laptops, and tablets. Thus, subject-independent, passive, and appearance-based models would hold a promising future owing to their wide applicability.

Automatic estimation of gaze direction is still challenging for remote methods [151]. This is due to the high variability presented in the eye image. Active methods reduce textural variations of the eye image, however, they have limited tolerance to occlusion and distance. For example, they are sensitive to external light reflections (e.g., caused by wearable glasses especially when associated with head movements) [151, 165]. It also demands very short distance between subject and light source to ensure sufficient illumination effect [166, 167]. In passive approaches, these lights are not used so they are more challenging than active methods.

For 2D images, some methods involve complex calibration setup in order to correctly estimate the gaze ray intersection with target plane. The target plane could be a screen or a wall with attached markers or even a virtual plane. The success of these techniques is dependent on the calibration process. For example, when calibrating using a computer screen, several gaze targets are displayed and subjects are requested to target their gaze on them in a sequence. Images of the eyes are captured for each point, and processed in order to find a mapping function between the screen coordinates
and these images. In depth-variant scenes, multiple recalibration may become necessary to solve inaccurate object depth estimation that cause parallax errors [168]. There are several approaches which enable algorithms to obtain such information (e.g., using multiple cameras, depth sensing devices, or modelling with 3D geometry).

Up till recent years, most reported algorithms have implemented an active approach which has achieved high PoG detection rates [169]; however, most reported EGT work in the literature, suffers from limitation to head movements and subject field of view, which make EGT less suitable for general purposes.

Various public gaze estimation datasets have been proposed in the literature, which can be used for training and testing EGT algorithms [163, 164, 170]. However, most EGT assessments are usually done within relevant method groups because fair comparisons of different algorithms are not always possible, due to the lack of standardisation as highlighted by recent reviews [151, 171]. Different dataset samples can have different image features (e.g., single eye or both eyes), annotations, labels and experiment settings that makes comparability a challenging task.

Instead, in our experiment we focus on 2D images of full-face annotated rich descriptive SSM parameters. In our case we used 17 facial features, including two reference points at the pupil’s centres. However, in our study we wish to highlight whether using SSM parameters can benefit in gaze estimation models. In our work we aim to compare several features such as geometrical-based and appearance-based features extracted from both eyes. We used a private dataset provided by the department, which contains 2D images of several subjects looking at different horizontal gaze angles in a car driving environment. Below we describe and discuss our experiments in more detail.
CHAPTER 5. GAZE ESTIMATION

5.6 Experiment

Our aim is to assess the correlation between eye gaze directions and parameters derived from different models such as the statistical shape models (SSM) and the appearance-based models. Our dataset contains 2D images of 10 subjects (7 males and 3 females) captured at 640 × 412 pixels. There were 7 gaze targets for each subject to look at spaced by 15°, one in the centre and three on each side as −45°, −30°, −15°, 0°, 15°, 30°, 45°. The gaze angle in this dataset was recorded in two ways, the angle relative to the centre front of the subject (where the camera is positioned at 0°) and the gaze angle relative to the head pose. For each head pose, subjects looked at all gaze targets within maximum range of 30° for both left and right directions. For every gaze and head-pose sample, three images were collected and they were annotated with 17 landmarks.

5.6.1 Feature Extraction

In order to train our model, we extract two geometrical features and appearance-based features. For the geometrical features, we built SSM model with 17 facial features as described in section 2.3, to extract the parameter $b$. Following the work of [158], we extracted the distance vectors between the eye’s corners and pupil’s centre. For the appearance features, a rectangular patch containing both eyes were extracted with two types of features: normalised grey intensities with global parameters that accounts for the position of the mid-point between the eyes in the image coordinate, the scale of the extracted patch and the rotation of the face in $y$ coordinate. For the second feature, we used haar-like features of integral images.

5.6.2 Regression Forest

We used a regression forest-based learning method with fixed parameters to examine the learning effect from different features. We split our dataset into: training set and test set. The training set contains 696 images of eight different subjects, while the test
Table 5.1: Statistical evaluation of gaze predictors with different input features

<table>
<thead>
<tr>
<th>Features</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaze intensities with global parameters</td>
<td>12.71</td>
<td>16.79</td>
</tr>
<tr>
<td>Integral of Haar-like</td>
<td>12.87</td>
<td>10.68</td>
</tr>
<tr>
<td>SSM parameters and Eye-corner distances</td>
<td>2.98</td>
<td>3.11</td>
</tr>
</tbody>
</table>

set contains 174 images of two subjects. we used all features to estimate the angle gaze relative to the head-pose. Figure 5.3 shows the cumulative distribution of absolute angular error for three different features: SSM parameters with eye-corners distances and two appearance-based features. The results show that features derived from geometrical SSM parameters and eye-corner distances have more potential than the appearance-based features alone. Statistical results are also shown in Table 5.1.

Figure 5.3: The cumulative distribution of gaze angular error of three different features.
5.7 Conclusion and Discussion

In this chapter we described general techniques used in eye gaze tracking, we also highlighted different challenges associated with different groups of techniques. Remote, passive and single camera-based approaches were the most challenging technique. Most EGT rely on a single eye image, while few works utilised information from both eyes [172]. Despite the large number of studies conducted in EGT, comparative evaluations are very challenging and mostly impractical because of the lack of standardisation.

In this chapter, recent state-of-the-art algorithms have been addressed (see summary section above). Most recent state-of-the-art algorithms adapt CNN-based approaches [150, 160]. CNN-based approaches have achieved great success not only in gaze estimation but many computer vision applications. We also performed a preliminary evaluation of several features such as SSM, eyes-corners distances and appearance-based features (integral images of Haar-like features) to assess their feasibility for EGT tasks under a regression framework. Our results suggest that geometrical features comprised of SSM and eyes-corners distances have more potential than appearance-based features alone.
Chapter 6

Parkinson’s Disease Case Study

6.1 Introduction

PD is a neurodegenerative disorder affecting around 10 million people worldwide [173] and about 1% of the population aged over 60 years [174]. An important symptom is the loss of muscle control leading to movement difficulties and effects on facial expressions.

Neurological Perspective

PD is associated with dopamine-deficiency due to the loss of dopaminergic neurons, the brain cells responsible for producing the dopamine which plays a vital role in essential human daily functions such as movement, memory, cognition, behaviour and attention. Patients with PD can suffer from motor disorders such as tremors, rigidity, slowness of movement (bradykinesia) and postural instability. Other non-motor symptoms linked to PD are disorders in mode, cognition and emotional processing [149, 175]. The disease can be directly linked to the loss of motor-control neurons in the substantia nigra. However it is still unclear whether this is because of the loss of non-motor control neurons in other parts of the brain or if it is just a side effect of the motor disorder...
Brain neurological regions such as the amygdala, basal ganglia, and right parietal cortices, are known to be involved in the recognition of various emotions [177]. This has led several studies to analyse emotional facial expression in brain-related neurological disorders such as Parkinson’s disease [175, 178] to make inferences regarding impaired emotions/expressions and their neurological-dependence [178].

**Diagnosis and Treatment**

There is no standard for diagnosing PD due to the lack of accurate measures [176, 179]. Many PD patients seek a neurological examination even though it is difficult for neurologists to verify the existence of PD due to the similarity of symptoms with those of other neurological disorders [176]. The best examination of PD to date is done by studying the dopamine system and the metabolism through a specialised brain scan, which is an expensive process [176].

Currently there is no cure for PD [179]. Several treatments exist which aim to reduce the symptoms such as physiotherapy, which provides general muscle exercises, occupational therapy such as training to perform specific tasks and speech and language therapy to encounter speech and sallowness problems caused by PD [179].

**Motivation and Goal**

Symptoms can vary significantly from one day to another. There are no quantitative measures of their severity, and so disease progression is usually evaluated by observation and interview with a clinical expert. This usually happens infrequently, and the results are not reliable because of the day-to-day variation in symptoms. This makes it difficult to manage treatment - for instance to evaluate whether the current drug regime is effective. An automatic method for measuring symptoms which could record them daily would enable much better management of the disease.
Currently many researchers are attempting to establish methods for assessing PD symptoms automatically using models which parametrise behavioural activities such as speech [180], emotional and non-emotional facial expressions [100] and emotional cognition processes [175, 181].

Facial expressions, in terms of face anatomical structure, are produced by movements of the facial muscles [182]. Therefore, accurate automatic characterization of facial expressions raises the potential for capturing the adverse effects of PD on those muscles. Despite the vast literature on automatic modelling of facial expressions, it is still an active area of research as most of the state-of-the-art methods are not robust or efficient enough for real-time applications [183].

The goal of this chapter is to explore the potential of feasible discriminative and quantitative measures for PD through modelling several emotional and non-emotional facial expressions. We examine measures of facial movement in both HC and patients with PD when making particular facial expressions. Our aim is to explore whether it is possible to measure the severity of symptoms related to facial muscles based on tracking the facial expressions of people with the disease.

6.2 Clinical Assessment in Parkinson’s Disease

In behavioural research, there is a debate on whether emotions are independent from cognition (see 4.3.1), therefore, previous analysis of PD from facial expressions spans over cognitive and facial-based motoric functions. In motoric facial impairments, researchers focus on facial masking effects and attempt to measure the severity of the disease on facial expressivity. On the other hand, research on cognitive impairments focuses on the perception of facial expression and is usually examined by a recognition task using prototypical categories of emotions [184, 185].
Cognitive Assessments

The cognitive impairment of PD can range from *Mild Cognitive Impairment* (MCI) to *Parkinson’s Disease Dementia* (PDD). The former type is associated with less severity such that it does not affect an individual’s daily activities; however, in the latter, the severity is extreme where it impacts daily activities. This type of impairment is known to be associated with reduced QoL and many social, financial and medical complications such as high stress on caregivers and demands for placement in nursing homes [186].

*Theory of Mind* (ToM) [187] is a concept which researchers use to evaluate social imagination and interaction, emotion recognition and understanding. ToM describes the individual’s ability to make inferences about their own mental state and that of others. These inferences are critical in human social interaction such that it can influence people judgments or actions [188]. ToM has been investigated in multiple disorders affecting cognition such as Alzheimer’s Disease (AD) [189, 190], Autism [191, 192] and PD. However, research studies regarding PD correlation with ToM abilities is still relatively recent [193] and heavily debated [194, 195].

The *Movement Disorder Society* (MDS) proposed guidelines for assessing impairments in PD known as *MDS revision of the Unified Parkinson’s Disease Rating Scale* (MDS-UPDRS) [196]. The MDS recommended the *Mini-Mental State Examination* (MMSE) [197], a 30-point scale measure, to be used for assessing cognitive functions in PD, which describes MMSE score of less than 26 points as prevalent of dementia for the younger population while the borderline for the older population is 24 points [186, 198]. Most PD studies rely on MMSE as a validation and screening tool for analysing non-demented patients.

Cognitive impairments have also been associated with other non-motor symptoms such as severe depression [199] and apathy [200], sleep disorder, pain, psychosis and genitourinary impairments [201, 202]. These associations, especially in later stages, makes
treating PD symptoms more difficult and often ineffective [202]. This has led clinicians to demand multiple screening tests for these non-motor symptoms to clearly understand PD progression and its contribution to the disability.

**Facial Expressivity Assessments**

There are several clinical guidelines/measures for evaluating and characterising PD. However, to the best of our knowledge, only two addressed facial masking in PD: the MDS-UPDRS [196] and *Interpersonal Communication Rating Protocol for Individual Expressive Behavior* (ICRP-IEB) [203].

As part of the MDS-UPDRS guidelines, in the second item of motor examination, (part III), facial expressivity is assessed by a rater’s general impression (gestalt impression) by an interview using a 5-point scale. However, the scale does not account for facial expression variations in terms of its configuration, regions, muscles or actions units (as defined by *Facial Action Coding System* (FACS)), which is an essential part for understanding emotional and pathological aspects of the disease [204].

The ICRP-IEB on the other hand, is designed primarily for evaluating PD expressive capacity from video recordings of standard social interactions with the subject undertaken at several points in time, typically at baseline, 6 weeks, 2 months and six months follow-ups. A systematic method is then used to extract short clips of the video (20 seconds) that best reflects the subject’s expressive state. These clips are then independently rated by multiple trained raters using a 5-point scale. The ICRP-IEB assesses facial expressivity in terms of actions such as eyebrow raising, blinking and cheek raising using rater’s gestalt impression for intensity, duration and frequency. However, ratings of some items such as *active mouth closure* require higher expert knowledge in speech and language pathology to enable sufficient reliability of the ratings [203].

Apart from general cognitive impairment and gestalt-based evaluations of facial masking, several behavioural studies investigated cognitive impairment related to specific
sets of facial expressions. Similarly, using multiple measurements of facial expressivity was suggested as a mean of assessing facial masking effects. These behavioural-based assessments of PD are discussed in more detail in the next section.

6.3 Facial Behaviour Research in Parkinson’s Disease

Measures of facial behaviour can be classified into two categories: message-based and sign-based [205]. The terms message-based and message judgment or sign-based and sign-judgment are used interchangeably in the literature [205–208]. In message-based, the measures are more holistic and used to describe the meaning behind the underlying expression (the message) using pre-defined categories or classes such as emotions, intentions, etc. In sign-based, the focus is to describe the configuration of the facial expression regardless of the message behind it. In previous analysis of PD, both measures have been used as the basis for evaluating facial expression impairments in PD.

In sign-based facial analysis of PD, some approaches aim to establish an objective quantitative measure for facial expressivity to aid clinical assessment of masking effects; while others focuses on qualitative measures that characterise PD impairments. For facial expressivity, the subject’s face is usually recorded or captured [10, 209, 210] to allow for multiple assessments of the subject’s facial expression intensity or quality, usually scored by a trained rater or by averaging scores of multiple raters[10, 209]. On the other hand, facial expression perception can be assessed using a classification task [175, 178] or from self-report measures that describe emotional feelings or experience [10].

The clinical support of these measures can be induced from correlation assessment with clinical measures of associated symptoms such as scores of depression [209], reduced QoL [211, 212], facial masking as in MDS-UPDRS [196] or as in ICRP-IEB [203].
However, in comparative studies involving HC samples, significant differences between HC and patients with PD can be highlighted using relevant statistical tests such as Analysis of Variance (ANOVA), Mann-Whitney or Wilcoxon.

6.3.1 Methods

There are several methods which can be used to analyse cognitive or expressive impairment. These can differ in terms of facial expression scope, stimulus type and experiment settings. Below we describe different data elicitation methods used in assessing facial expression impairment in both cognition and production.

Stimulus Representation

Before we discuss different types of stimuli, it is essential to understand that facial expressions can be classified into three categories: spontaneous, self-posed and imitated expressions. These can also be classified in terms of emotional and non-emotional expressions. Based on the nature of the facial expression and the assessment task (expressivity or cognition), certain methods can be used to elicit facial expression responses.

Stimuli used in facial expression research can be classified into visuals (static image or video), verbal or written. Visual stimuli that are used to elicit facial expression responses can represent real faces, artificially altered (e.g., using AAM) or created (e.g., using cartoons or avatars). Written or verbal stimuli can represent emotional statements or labels of emotions or expression categories.

Facial Expression Recognition in PD

There are a substantial number of studies addressing FER impairment in PD. Here, we discuss some of the most relevant work that has been done in this area to provide some insight to variables such as goal, stimuli type, nature of facial expression, subject group, etc., that play important roles in study design and outcomes.
In cognitive studies, stimuli can be used to analyse cognitive impairment related to brain-specific modality. This can help researchers to understand how the brain processes facial expressions induced from different stimuli. For example, to make inferences regarding related neurological substrate as whether the neurological contribution of emotional perception from different stimuli are shared or separate [175, 178].

For example, Kan et al. [178] compared emotion recognition from three types of stimuli _facial expression, prosodic_ and _written statements_, when comparing 24 HC and 16 subjects with PD. All stimuli from every modality were designed to represent one of the six basic emotions, and subjects were asked to select emotional labels, which best described the emotion state of the stimuli using pre-labelled cards. Facial expression stimuli are presented from video recordings of actors performing a target expression while prosodic stimuli represent audio recordings of actors reading emotional statements. Impairments were only seen in facial expression recognition with PD, particularly in the recognition of fear and disgust. A standard assessment of depression along with several cognitive functions such as memory, attention, reasoning and planning were also collected in order to examine the correlation between each test. Positive correlations with facial expression modality were found between fear and scores of depression.

Sprengelmeyer et al. [175] also examined PD defects on the perception of facial expressions in terms of emotions, sex, age, familiar faces and unfamiliar faces using standardised tests. They compared 40 HC to 36 subjects with PD at perceiving the six basic emotions using prototypical and morphed images as in (_Ekman 60 Faces test_ and _Emotion Hexagon_) in _Expressions of Emotion: Stimuli and Tests_ (FEEST) [213] respectively. The _Ekman 60 Faces_ contains six prototypical images of 10 subjects representing the six basic emotions while the _Emotion Hexagon_ contains morphed images of the basic emotion produced from [214]. Differences were examined on all tests using two-way repeated measures ANOVA. Significant differences were only seen in the recognition of disgust and anger facial expressions.
Measures derived from FER in PD are mainly message-based where the primary interpretation of facial expression is performed by the subject. The primary assessment in cognitive studies of PD usually implies a classification task to describe emotional experience or recognition problems in PD. However, emotional experience of PD patients has been found to be relatively unaffected [10]. On the contrary, while it is well accepted that emotional recognition is impaired in PD, studies that aim to identify impaired emotional FER are still contradictory. A recent review published on 59 emotional recognition studies shows contradictory findings of impaired emotional FER in PD, and highlights the need for quantifiable measures to address neurological loss in PD [215]. While there is a substantial literature on emotional facial recognition in PD little work has been done in quantifying facial expressivity in PD.

### 6.3.2 Facial Expressivity Assessment in PD

In expressivity analysis, certain stimuli can be used to elicit a certain type of facial expression. Posed expressions are produced often with verbal or written labels describing the emotion/expression [210]; while imitated facial expressions are produced with visual aids by displaying typical examples of the expression. This can be done by face-to-face training, watching video clips [10, 216] or presenting static images [217]. Spontaneous expression on the other hand, are evoked either from watching video clips [10, 209] or during interactive discussions or interviews [10, 211, 218]. Facial expressivity under regulated discussions or interviews are measured in terms of valence, regardless of emotion category or class [10, 203].

Apart from quantifying facial expressivity in different emotions or expressions, or identifying their abnormalities, some studies aim to analyse facial expressivity as markers of emotional experience [219, 220] or assess patients ability to voluntarily mask emotional expressions (incongruent expressions). For example, when posing a smile while watching an amusing video clip [10] or when smelling unpleasant odours [221]. These studies are considered out of the scope of this project.
Both message- and sign-based methods can be used to analyse facial expressions in PD. However, message-based measures are derived from signs and cues presented on the face and are more suited for the purpose of AMFE. Below we discuss previous sign-based attempts to analyse facial expressivity in PD in terms of its encoding method, facial expression type, studied subjects, task (recognition/regression), sampling method (manual/automatic), experimental settings, analysis method and research outcome in relation to PD impairments on facial expressions.

### 6.3.3 Past studies

One of the earliest attempts to quantify PD effects on facial expression was conducted in 1988 by Katsikitis and Pilowsky [209]. A total of 18 participants (9 HC and 9 with PD) were studied in order to identify significant differences of facial muscle movements between the two groups during smiling experience. Subjects were recorded while viewing 11 video slides, which were used as a stimuli to elicit smiles. Video frames of recorded subjects were then manually filtered in order to select the best animated smile of each subject then digitised using 62 facial landmarks in a still photographs using a standard procedure [222]. Each photograph was scored based on a set of horizontal and vertical distances between several facial features (12 measures in total). All sided measures were averaged with their counterparts and horizontal measures were normalised by the horizontal distance of the external eyes corners, while vertical distances were normalised by the height of the nose. The ground truth of smiling frequencies were rated by two independent raters. Their findings suggested that smiling frequency that is associated with mouth opening was significantly higher in the control group.

The quantification of facial expressivity described in [209] did not account for the dynamics of facial expressivity as each rater’s scores were based on a single digitised image for each subject. The approach is also not suitable for the purpose of automatic estimation of facial expression intensity as the process of obtaining these images involves
6.3. FACIAL BEHAVIOUR RESEARCH IN PARKINSON’S DISEASE

manual filtering and digitisation of facial landmarks. This is probably due to insufficient automatic facial feature tracking at that time [71]. Furthermore, only geometric features of spontaneous smile expression were analysed for PD defects.

Simons et al. [10] compared facial expressivity in 19 people with PD to 26 HC from video recordings, which contains participant’s facial reactions to different stimuli. Participants undertook several tasks such as watching amusing video clips and short discussions of enjoyable topics, each in multiple different contexts, with familiar and unfamiliar people. Participants were also asked to pose the six basic emotions, mask spontaneous smile with disgust, when watching amusing video clips, and imitate nine non-emotional facial movements based on FACS. For each recorded video, a short segment was manually selected, which showed the most expressive behaviour of the subject. Subjective (self-reports) and objective (FACS) measures were used to assess facial expressivity. In every situation, facial expressivity was scored by a trained rater and derived from participant’s self-report based on 100-point likert scale. Correlations between both ratings revealed reduced facial expressivity in general. However, within the PD group, there was a significant increase in expressivity level for spontaneous facial expression in cases involving an experimenter or a stranger.

This study showed how social context could significantly influence the expressivity level of spontaneous facial expression for people with PD; however, there are some limitations to the study. As suggested by the author, the presence of an attentive and unfamiliar person could enhance emotional experience, thus, increasing the expressivity level in spontaneous expressions, while this was evident in the spontaneous expressions, this assumption remains unexplored in posed expression. Moreover, unlike posed expressions, spontaneous expressions were limited to smiles, therefore, comparison to other spontaneous emotions was not feasible.

Another work presented by Bowers et al. [210] quantified facial dynamics from videos based on an entropy, a measure originally developed from [223], which relates to the
mean change in pixel intensity between two successive frames. They used a semi-
automatic method to track 20 facial landmarks from video recordings of subjects while
presenting several voluntary expressions. A total of 24 subjects were involved in this
study (12 with PD and 12 HC). Based on this entropy, two measures used to compute
the onset till entropy’s peak: the total movements change and the latency. Compared
to HC, PD subjects showed less total movements on all emotions; however, it is unclear
whether some emotions were more impaired than others. For emotion comparison,
their findings suggest that surprise and anger presented significantly more movements
change, while sad has the least. The overall latency was measured between the two
groups and for each emotion independently. The latency was greater in PD, while
amongst emotions, fear was the lowest.

The method used in [210] is not suited for the purpose of automatic daily monitoring
as it is highly constrained. The subject’s head was restrained on a special device
over a normalised lighting rig to eliminate out-of-plane head movements. The onset
detection is also not computed directly from the expression data, instead the onset is
estimated from an attached buzzer which instructs the subject to produce the target
expression.

Another sign-based approach that adopted the FACS scheme to quantify facial ex-
pressivity was presented by Wu et al. [216]. In their approach, a context-independent
action unit recognition system [224] along with automatic facial feature tracking [225]
was used. Both systems were used to extract relevant features from video frames
of subjects recorded while watching emotion-stimuli video clips. Geometry-based and
appearance-based features were extracted by the tracking system while the AU recogni-
tion system was used to label active action units from a set of 11 action units. For every
action unit a binary classifier was built using a Support Vector Machine (SVM). The
input vector was derived from a combination of geometry-based and appearance-based
features, which were extracted from tracking 83 facial features. To reduce dimen-
sionality, a feature selection was applied on the input vector using an adaptive boosting
algorithm. The output vector was a set of labels generated by the action-unit recognition system defining the associated active AUs in each frame. Expression intensities were estimated from segments of 30 second intervals using a probability function proposed in [226], which is defined by a sigmoid function that incorporates the overall SVM’s marginal distances of relevant active AU. Two measures were induced: the displayed action intensity per segment and the total amount of movement change from a neutral baseline.

The above method combines sign-based and message-based measures. Message-based measures, self-reports, were used to narrow the assessment to significantly felt emotions using ANOVA with repeated measures. Sign-based analysis was applied to disgust facial expressions. Despite the reported significance, the system may not be reliable under naturalistic settings. The action-unit recognition system was trained and tested on the Kanade dataset which contains images of subjects taken under constrained laboratory settings. Similarly, in experiment settings, subjects’ samples were collected using an invasive and constrained method. While it is true that disgust-induced video clips gave the most significantly "felt" emotion, it is not necessarily evidence that PD impairments are not correlated with different emotions.

Vinokurov et al. [227] compared 14 PD patient with 15 HC while watching several amusing clips. Appearance facial features were encoded using FACS, which were extracted using Faceshift, a commercial software that utilises a 3D depth camera to track 51 AUs. Statistical measures such as mean, variance, skew and kurtosis were derived from every AU and their bilateral differences. Facial masking was assessed using MDS-UPDRS scale, rated by two independent raters. Greedy Linear Regression (GLR) learning with Leave-one-out method was used to establish two independent models, each built from single rater scores. Correlations between model scores and both raters were examined and found to be significant.

Despite this significance, several questions remain unanswered such as which facial expressions or AU are impaired and to what degree. The method described requires
calibration steps for every subject. It also requires a 3D depth camera, which is not as widely accessible as traditional 2D cameras.

Joshi et al. [211] used supervised learning Random Forest Regression (RFR) to estimate QoL scores of PD subjects. A total of 727 20-seconds video clips were collected from the Self-Management Rehabilitation for PD database [228], which contains video recordings of 117 PD patients during social interaction with a clinician. Video clips were rated by four experts using the 5-point Likert scale to describe QoL scores. The mean score was used as the ground truth score for each video. The input feature was formulated by combining quartiles and several statistics derived from the first derivative of a set of normalised horizontal and vertical Euclidean distances (6 in total), which were extracted from tracking 59 facial features. For each distance, four quartiles, the standard deviation and total local maxima per unit were encoded to represent the input feature. RFR with 9-fold cross validation was used. Feature importance was examined using out-of-bag error estimation [49], which evaluates prediction errors of multiple features using random permutations to training samples that are not used in the construction of the tree. The differences in the average mean square error (MSE) were used to highlight feature importance. The results in the paper suggest that movements related to eyes and eyebrows are better indicators for QoL scores than the movements of the mouth.

In tracking video frames, it was mentioned that there was a significant number of untracked frames but there was no quantifiable information for this loss, it would be interesting to understand how much information would be required to achieve such performance. Evaluation of input features was also random over temporal space and did not explore defined dynamics of facial expression such as the onset and the offset as they have been found to be distinguishing features in many behavioural studies [229]. Since only PD subjects were studied in this work, no comparison was available for highlighting normal and abnormal behaviour. For this reason, the analysed features were limited to those that best describe the scores of QoL selected by RFR model and
6.3. FACIAL BEHAVIOUR RESEARCH IN PARKINSON’S DISEASE

did not explore other possible differences between HC and subjects with PD.

Bandini et al. [217] analysed facial expressions in PD when imitating or acting of four facial expressions: anger, disgust, happy and sad. Seventeen HC were compared to 17 with PD in this study. For every subject, there were 9 recorded videos: 1 for neutral and 2 videos (acted and imitated) for each expression. An automatic facial feature tracker [230] was used to extract 49 facial landmarks from these videos. For every subject, a mean shape (neutral baseline) was produced from only neutral video frames by averaging shapes as in the PCA method. The expressivity of the face was estimated as the residuals of 20 Euclidean distances from the neutral baseline, after mapping the tracked features to the reference frame (the mean shape) using an affine transformation with 4 reference points (inner eyes’ corners, the mid point between them and the nose tip). Five statistical measures of expressivity were computed: the mean, the standard deviation, the maximum, the minimum value and the range. Significant differences between the two groups were found in the imitation of anger, disgust, happy and sad however, for acted expressions significant differences were only seen in disgust, happy and sad. Unlike HC, PD did not show significant differences between acted and imitated expressions.

Bandini et al. [217] also assessed subjects’ presentation of facial expressions using the results of an automatic multi-class facial expression recognition as a benchmark. They trained multiple SVM classifiers, one for each expression and the test samples were labelled based on the classifier with the highest confidence. Their findings suggest that HC subjects produce less ambiguity in their facial expressions when compared to PD subjects.

It is generally accepted that PD subjects suffer from facial masking effects as well as impairments in cognitive functions. For example, subjects may tend to imitate a certain expression but with the wrong emotion. It may be necessary before comparing subject’s ability to imitate certain expressions to assess their cognitive functions to ensure that they perceived the right emotion, using standardised cognitive assessment
such as the MMSE. This step was not addressed in [217]. Another limitation is that
the multi-classifier was not reliable for emotion recognition, especially as reported in
the recognition of happy expressions. First, the training samples were too small for
building generic models. The 840 frames collected for happy expressions represent
the last 4 frames of each video which would be expected to show little variation.
Second, in testing, the classifier was used to classify frames depicting the dynamics of
emotional expressions from the complete frame sequence, while it was not trained to
do so. Instead, it was trained to recognise the emotion only from the last four frames
where the expressivity level is at its peak, therefore, it did not learn the expression’s
course.

Joshi et al. [218] explored the effects of two contextual variables: gender and sentiment
on facial expressivity predictions. They used the Self-Management Rehabilitation of
PD dataset[228], which contains a 805 short video clips, $\approx$ 20 seconds each, recorded
from 117 subjects. Facial expressivity was rated by 4 trained raters based on ICRP-IEB
[203] items of active expressivity defined in 5-point likert scale, the mean score was used
as the ground truth for the expressivity rating. Appearance features defined by FACS
were successfully extracted from 772 videos using OpenFace [231], while audio fea-
tures Mel-Frequency Cepstral Coefficient (MFCC) were extracted using Librosa library
[232]. For each frame, scores of AUs intensities and their activations were recorded
along with the audio features of MFCC. For each video, four statistical measures were
computed for every visual and audio feature: mean, standard deviation, minimum
and maximum. They constructed several single and multimodal Hierarchical Bayesian
Neural Networks (HBNN) to classify and estimate expressivity ratings in PD. With
9-fold cross validation, they compared several HBNNs with and without contextual in-
formation to predict facial expressivity, for both regression and classification problems.
Evaluation of context-sensitive HBNN, revealed that only sentiment-based contextual
information improved expressivity prediction in both tasks. Feature importance was
only estimated for visual features by averaging the importance of each of the derived
statistical measures related to the presence and the intensity of AUs.

6.3.4 Summary

PD’s influence on emotional facial expressions remains an active area of research with a number of unresolved challenges [178, 215, 219, 233–236]. First, contradictory findings persist regarding which emotional expressions are impaired in PD and to what extent [181, 234, 236]. Moreover, while most PD studies have focused on emotional FER, very few attempts have been made to examine facial display [215]. Tables 6.1 and 6.2 provide a summary of these methods in terms of the analysed sample and approach.

Although several attempts have been made to assess the dynamics of facial expression in PD, very few have established an automatic approach, and quantitative measures have not always been used [217]. Furthermore, most of these quantifications are based on discrete measures, such as FACS, which does not describe the full range of facial behaviour.

A limited number of automatic methods have been proposed to assess the PD facial-mask effect, and they are associated with several drawbacks. First, none were intended for the purpose of daily monitoring in naturalistic settings. Second, all automatic assessments of facial expressivity were implicit, using global statistical measures over the course of the expression [211, 216, 227]. These methods lack explicit measures that define temporal variations of facial dynamics, such as the onset, apex and offset, which are crucial cues in the emotional processing of facial expressions [237].

Furthermore, authentic spontaneous expressions are difficult to elicit, especially when subjects are fully aware of the experimental setting [71]. Although all automatic PD assessments were performed in laboratory settings regardless of the facial expression’s nature, some studies analysed spontaneous expressions with invasive methods [216], thus rendering their authenticity vulnerable.

Another limitation in previous methods of automatic PD analysis is the range of facial
### Table 6.1: Description of PD case studies’ dataset.

<table>
<thead>
<tr>
<th>Papers</th>
<th>Expression nature</th>
<th>No. of Diff. Expressions</th>
<th>Samples per expression</th>
<th>Samples per subject</th>
<th>Age (µ, ± SD)/range</th>
<th>Medicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katsikitis and Pilowsky [209]</td>
<td>S, P, M, S &amp; SSI</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(69.2, [60,77])</td>
<td>Y</td>
</tr>
<tr>
<td>Simons et al. [10]</td>
<td>smile BE, FACS actions, smiles &amp; AEBE</td>
<td>2</td>
<td>smile BE, FACS actions, smiles &amp; AEBE</td>
<td>smile BE, FACS actions, smiles &amp; AEBE</td>
<td>(67.37 ± 8.42)</td>
<td>Y</td>
</tr>
<tr>
<td>Bowers et al. [210]</td>
<td>NE &amp; BE</td>
<td>3</td>
<td>1 still image</td>
<td>1 still image</td>
<td>(65.6 ± 9.1)</td>
<td>Y</td>
</tr>
<tr>
<td>Wu et al. [216]</td>
<td>S, N.A</td>
<td>1</td>
<td>1 still image</td>
<td>1 still image</td>
<td>(N.A, [47,76])</td>
<td>Y</td>
</tr>
<tr>
<td>Vinokurov et al. [227]</td>
<td>P, M &amp; SSS</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(N.A, [48,84])</td>
<td>Y</td>
</tr>
<tr>
<td>Joshi et al. [211]</td>
<td>ISO</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(68.9 ± 7.5)</td>
<td>Y</td>
</tr>
<tr>
<td>&amp; Joshi et al. [218]</td>
<td>ISO</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(71.9 ± 9.2)</td>
<td>Y</td>
</tr>
<tr>
<td>Bandini et al. [217]</td>
<td>ISO</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(66.3 ± 9.0)</td>
<td>Y</td>
</tr>
<tr>
<td>&amp; Mbalik et al. [205]</td>
<td>ISO</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(71.9 ± 9.2)</td>
<td>Y</td>
</tr>
<tr>
<td>&amp; Mbalik et al. [206]</td>
<td>ISO</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(66.3 ± 9.0)</td>
<td>Y</td>
</tr>
<tr>
<td>&amp; Mbalik et al. [207]</td>
<td>ISO</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(71.9 ± 9.2)</td>
<td>Y</td>
</tr>
<tr>
<td>&amp; Mbalik et al. [208]</td>
<td>ISO</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(66.3 ± 9.0)</td>
<td>Y</td>
</tr>
<tr>
<td>&amp; Mbalik et al. [209]</td>
<td>ISO</td>
<td>1</td>
<td>1 video</td>
<td>4</td>
<td>(71.9 ± 9.2)</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Expression nature**
- ISO: Standardised social interaction
- S: spontaneous
- P: posed
- M: mimicked

**Class**
- NE: neutral
- BE: Basic Emotions
- AE: Active Expressivity

**No. of Diff. Expressions**
- Number of different expressions

**Expression nature**
- SSI: Standardised social interaction
- S: spontaneous
- P: posed
- M: mimicked

**Expression nature**
- ISO: Standardised social interaction
- S: spontaneous
- P: posed
- M: mimicked

---

Papers: Katsikitis and Pilowsky [209], Simons et al. [10], Bowers et al. [210], Wu et al. [216], Vinokurov et al. [227], Joshi et al. [211], Joshi et al. [218], Bandini et al. [217], & Mbalik et al. [205].

Table 6.1: Description of PD case studies’ dataset. The order of items in the table corresponds to the order in the text.
Table 6.2: Summary of PD case studies analysis methods

<table>
<thead>
<tr>
<th>Papers</th>
<th>Auto. Tracking (tool)</th>
<th>Auto. Learning</th>
<th>No. of Facial Landmarks</th>
<th>Quantification</th>
<th>Input Feature</th>
<th>Features Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katsikitis and Pilowsky [209]</td>
<td></td>
<td>62</td>
<td>✓</td>
<td>Static Dynamics</td>
<td>Geometric</td>
<td>12 relative distances</td>
</tr>
<tr>
<td>Simons et al. [10]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bowers et al. [210]</td>
<td>semi</td>
<td>20</td>
<td>✓</td>
<td>Appearance</td>
<td></td>
<td>Entropy [223]</td>
</tr>
<tr>
<td>Wu et al. [216]</td>
<td>✓</td>
<td>SVM</td>
<td>83</td>
<td>Geometric &amp; Appearance</td>
<td></td>
<td>a shape model of 11 Action Units</td>
</tr>
<tr>
<td>Vinokurov et al. [227]</td>
<td>semi (Faceshift)</td>
<td>GLR</td>
<td>N.A</td>
<td>Appearance</td>
<td></td>
<td>51 Action Units</td>
</tr>
<tr>
<td>Joshi et al. [211]</td>
<td>✓</td>
<td>RFR</td>
<td>59</td>
<td>Geometric</td>
<td></td>
<td>6 statistical quartiles of Euclidean distances</td>
</tr>
<tr>
<td>Bandini et al. [217]</td>
<td>✓</td>
<td>SVM</td>
<td>49</td>
<td>geometric</td>
<td></td>
<td>20 (Angles &amp; Euclidean distances)</td>
</tr>
<tr>
<td>Joshi et al. [218]</td>
<td>✓ (OpenFace)</td>
<td>HBNN</td>
<td>59</td>
<td>Appearance &amp; 35 AUs (activation/intensities) &amp; 4 MFCC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
expressions explored is not as wide as in manual methods (see Table 6.1). Small sample sizes are a common problem in most clinical studies, which makes predictive models more sensitive to overfitting unless the way features are selected is controlled (e.g., using LASSO). None of the previous automatic methods have examined this topic.

6.4 The Case Study

Thirty-one participants (24 HC and 7 with PD) were recruited by specialised experts to facilitate multiple analytical studies of PD impairments, including cognition, speech, finger tapping and facial masking. Several screening tests were conducted to evaluate several perceptual and motor functions, such as the The Netherlands Optical Society (TNO) stereoacuity test [238], MMSE, Geriatric Depression Scale (GDS) test [239] and Minimum Data Set (MDS) [240]. All participants analysed in this study were classified non-depressive with normal cognitive function. A summary of the demographics and two screening tests for both groups are shown in Table 6.3.

Table 6.3: Summary of subjects demographics and screening tests

<table>
<thead>
<tr>
<th>Demographics &amp; screening tests</th>
<th>HC</th>
<th>PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Mean: 68.1</td>
<td>Mean: 56.7</td>
</tr>
<tr>
<td></td>
<td>SD: 1.1</td>
<td>SD: 6.8</td>
</tr>
<tr>
<td></td>
<td>Min: 67.3</td>
<td>Min: 56.3</td>
</tr>
<tr>
<td></td>
<td>Max: 69.7</td>
<td>Max: 72.4</td>
</tr>
<tr>
<td>Years of Education</td>
<td>Mean: 19</td>
<td>Mean: 14.86</td>
</tr>
<tr>
<td></td>
<td>SD: 0.58</td>
<td>SD: 3.48</td>
</tr>
<tr>
<td></td>
<td>Min: 18</td>
<td>Min: 8</td>
</tr>
<tr>
<td></td>
<td>Max: 19</td>
<td>Max: 18</td>
</tr>
<tr>
<td>MMSE test</td>
<td>Mean: 29.75</td>
<td>Mean: 29</td>
</tr>
<tr>
<td></td>
<td>SD: 0.5</td>
<td>SD: 1.15</td>
</tr>
<tr>
<td></td>
<td>Min: 29</td>
<td>Min: 27</td>
</tr>
<tr>
<td></td>
<td>Max: 30</td>
<td>Max: 30</td>
</tr>
<tr>
<td>GDS test</td>
<td>Mean: 0.25</td>
<td>Mean: 2.57</td>
</tr>
<tr>
<td></td>
<td>SD: 0.5</td>
<td>SD: 2.99</td>
</tr>
<tr>
<td></td>
<td>Min: 0</td>
<td>Min: 0</td>
</tr>
<tr>
<td></td>
<td>Max: 1</td>
<td>Max: 9</td>
</tr>
</tbody>
</table>

The HC group consists of two sub-groups: (i) PD-age matched (N=4) subjects were between the ages of 67.3 and 69.7 years with mean=68.1 and SD=1.1 and (ii) 20 subjects between the ages of 20 and 47 with mean=29.9 and SD=6.1 years. The PD group subjects were between the ages of 56.3 and 72.4 years (mean=56.7, SD=6.8) and were all diagnosed with the early stages of PD, between 0 and 2 as defined by Hoehn and Yahr scale as part of an MDS-UPDRS assessment (part IV motor complications).
All PD patients began taking medication (Levodopa) prior to the study, except for one participant who began taking medication half-way through the testing period.

6.4.1 Data collection

All the subjects’ data were collected over periods of varying duration. Once per day, they used a computer program that prompted them to perform various facial expressions as described in section 4.5. All participants were instructed to perform three expression sequences: posed, mimicked and non-emotional actions as in Figure 4.5. Participants were recorded on a daily basis, one week for the PD-age matched group and a single day for the remaining 20 participants. However, subjects with PD were followed for six weeks.

Data collection was conducted in two trials. In the first trial, all participants performed a single instance of each unique expression every day. In the second trial, the 20 participants performed similar sequences, however, with a varying number of repetitions for each sequence, spaced randomly within the sequence group. A detailed summary of the number of expression samples gathered for each group are presented in 6.4, while Figure 6.1 depicts a summary of the data collection process.

Table 6.4: Dataset summary showing total numbers of displayed expressions for each group

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Subjects</th>
<th>Total weeks</th>
<th>Days per week</th>
<th>Expressions per day</th>
<th>Repetitions</th>
<th>Total samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC</td>
<td>20</td>
<td>NA</td>
<td>1</td>
<td>24</td>
<td>3</td>
<td>1,440</td>
</tr>
<tr>
<td>HC</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>18</td>
<td>1</td>
<td>1,800</td>
</tr>
<tr>
<td>PD</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>18</td>
<td>1</td>
<td>3,780</td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>7</td>
<td>11</td>
<td>54</td>
<td>5</td>
<td>7,020</td>
</tr>
</tbody>
</table>
At the end of each session, a set of questions designed by specialists was displayed on the screen to assess the patient’s QoL. Subjects responded verbally to these questions while being recorded by the laptop computer’s microphone and camera. These answers were then evaluated by five independent raters using scores ranging from 1 to 9, with 1 being the worst case and 9 describing the best case. The average score was used as the ground truth in this study.
6.4.2 Annotation

In the initial trial, transitions between different expressions' stimuli were fully manual, managed by participants via mouse clicks; hence, the annotation of participants' responses was also manual. However, to simplify the annotation task, in the second trial, manual transitions were only enabled between different sequences (e.g., between self-posed and mimicked); more specifically, participants had to self-time themselves between sequences. Once a sequence of stimuli began, each stimulus was displayed at known time intervals, followed by a three-second countdown during which participants were expected to display their facial expression response. Therefore, manual annotations were only performed for the beginning of the sequence, while the remaining responses were automatically predicted. All expressions were then annotated with metadata indicating the beginning and the end of the expression development. These predictions were used as initial estimates and further refined as will be described later in section 6.4.4.

6.4.3 Tracking Facial Features

Because of the unconstrained nature of the recording, facial feature detection and tracking are challenging. To improve tracking results, we adopted two facial feature tracking techniques: a generic RFCLM [22] and subject-specific AAMs as described in Chapter 3. Subject-specific AAMs were used for extremely challenging conditions as they are more robust and easier to build and require fewer training examples than generic models.

A 27- and a 51-point-based tracker were used to extract features from video frames within the expression’s time interval. To evaluate the tracking performance while accounting for variance bias caused by different frame rates and times, we computed the rate of successfully tracked frames for each session independently. Then, we computed their mean and standard deviations for every subject and expression; the results are
shown in Figure 6.2.

Figure 6.2: Mean and standard deviations of tracked frames rates per session. First and second rows represent, the 27-point and 51-point, tracking results respectively, while the first and second columns show categorised results per expression and subject, respectively.

The Figure in 6.2 highlights that the largest loss of frames occurred for participant SC107 ($\mu=0.58$) when tracked with the 27-point tracker. It also shows that the 51-point tracker is generally more robust than the 27-point tracker. This suggests that
facial feature tracking under naturalistic settings is still challenging; however, training trackers with more facial landmarks can improve tracking performance. In the next section, we examine the tracked data to quantify and qualify them for further analysis as explained below.

6.4.4 Irrelevant Expression Data Removal

To enhance model learning, all frames that were collected from initialising a tracker on a false region or with insufficient expression samples due to significant tracking loss were manually filtered from the HC dataset. However, in the second trial, due to the large size of the dataset, cleaning the data manually became impractical and excessively time-consuming. Therefore, we applied an automatic method to detect and quantify these tracking failures.

First, we computed the geometric features $\delta B$ (see section 4.6.1, Chapter 4). Since the systematic annotation of an expression’s onset and offset are an approximation, it can potentially include large neutral frames, leading to the neutral-biased effect. Therefore, we reduced the neutral samples by examining their relative distribution in all the other expressions’ subspace (see Section 4.6.4 in Chapter 4). After reducing the neutral samples, short-tracked segments that are insufficient for behavioural analysis were also discarded. Next, we improved annotations of the expressions’ onset and offset by following their gradient over time (see section 4.6.5, Chapter 4). Finally, we quantified the rates of these tracking failures, manifested by tracking over a false region (i.e., false positives) or due to short segments (see Section 4.6.6 in Chapter 4).
Figure 6.3: Quantified 27-point tracking failures manifested in false tracking and short segments. Figures (a) and (b) shows percentages of discarded frames per participant and expression.
6.4. THE CASE STUDY

Figure 6.4: Quantified 51-point tracking failures manifested in false tracking and short segments. Figures (a) and (b) shows percentages of discarded frames per participant and expression.
Figures 6.3 and 6.4 show rates of estimated tracking failures for both the 27- and 51-point-trackers, respectively. In each figure, results are shown per subject (lines) and expression (bars). These results reveal that the expressions of blinking, mimicked disgust and fear were associated with the most tracking failures, between approximately 9.78% and 14% of the total frames for the 27-point shapes and around 4% for fear, anger and mimicked disgust for the 51-point shapes. Furthermore, the 27-point geometrical features of subject SC107 were associated with the most tracking failures, suggesting that facial tracking for this subject was challenging with the 27-point tracker. As we examined the recordings of SC107 we found that the camera was exposed to extremely bright light through a wide window behind the participant, which sometimes rendered the facial region mostly invisible even to the human eye. The 27-point tracker was more vulnerable to this effect than the 51-point tracker as demonstrated by Figures 6.3 and 6.4.

6.4.5 Geometrical shape analysis

Next, we model the expression subspace based on the variation of parameter \( c \) (described in section 4.6.2, Chapter 4) from two statistical shape models, a 27- and a 51-point model.

Since displayed expressions vary over time and to analyse expression curve variations, we normalised each displayed expression to an equal length by linear interpolation in time. Then, we modelled their variations using the PCA method. Table 6.9 shows the effects of varying several modes within \( \pm 3 \) standard deviations about the neutral point. Note that the direction of facial expressivity of different modes is not always in a negative direction. An example can be seen in mode 0 for happy expressions in the 51-point-based model. After this process was completed, each displayed expression was composed of 50 frames.

Tables 6.5 to 6.8 provide an overview of the size of our dataset pre- and post-normalisation.
for both HC and PD groups along with details about the total number of sessions, duration, total number of frames and the total resampled frames. The final sampling aimed to produce comparable instances of facial expression with a reduced neutral-biased effect as described in section 4.6.4 in Chapter 4.
Table 6.5: The table shows the number of original frames extracted from video sequences of different control subjects using two facial feature trackers. The *Frames (re-sampled)* column describes the total number of re-sampled frames after reducing neutral frames and refining annotations of onset and offset.

<table>
<thead>
<tr>
<th>Sessions Duration (in seconds)</th>
<th>Total Frames</th>
<th>re-sampled Frames</th>
<th>51-point tracking data</th>
<th>27-point tracking data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject 2</td>
<td>736</td>
<td>297</td>
<td>66</td>
<td>1235</td>
</tr>
<tr>
<td>Subject 3</td>
<td>746</td>
<td>36</td>
<td>88</td>
<td>1000</td>
</tr>
<tr>
<td>Subject 4</td>
<td>782</td>
<td>62</td>
<td>90</td>
<td>906</td>
</tr>
<tr>
<td>Subject 5</td>
<td>868</td>
<td>187</td>
<td>06</td>
<td>109</td>
</tr>
<tr>
<td>Subject 6</td>
<td>564</td>
<td>195</td>
<td>18</td>
<td>185</td>
</tr>
<tr>
<td>Subject 7</td>
<td>450</td>
<td>61</td>
<td>90</td>
<td>101</td>
</tr>
<tr>
<td>Subject 8</td>
<td>474</td>
<td>114</td>
<td>25</td>
<td>110</td>
</tr>
<tr>
<td>Subject 9</td>
<td>450</td>
<td>108</td>
<td>07</td>
<td>109</td>
</tr>
<tr>
<td>Subject 10</td>
<td>474</td>
<td>113</td>
<td>07</td>
<td>109</td>
</tr>
<tr>
<td>Subject 11</td>
<td>502</td>
<td>119</td>
<td>07</td>
<td>110</td>
</tr>
<tr>
<td>Subject 12</td>
<td>416</td>
<td>135</td>
<td>07</td>
<td>120</td>
</tr>
<tr>
<td>Subject 13</td>
<td>474</td>
<td>113</td>
<td>07</td>
<td>109</td>
</tr>
<tr>
<td>Subject 14</td>
<td>502</td>
<td>119</td>
<td>07</td>
<td>110</td>
</tr>
<tr>
<td>Subject 15</td>
<td>534</td>
<td>138</td>
<td>07</td>
<td>120</td>
</tr>
<tr>
<td>Subject 16</td>
<td>474</td>
<td>113</td>
<td>07</td>
<td>109</td>
</tr>
<tr>
<td>Subject 17</td>
<td>502</td>
<td>119</td>
<td>07</td>
<td>110</td>
</tr>
</tbody>
</table>

The table shows the number of original frames extracted from video sequences of different control subjects using two facial feature trackers. The *Frames (re-sampled)* column describes the total number of re-sampled frames after reducing neutral frames and refining annotations of onset and offset.
Table 6.6: The table shows the number of original frames extracted from video sequences from control dataset (by expression) using two facial feature trackers. The Frames (Re-sampled) describes the total number of re-sampled frames after reducing neutral frames and refining annotations of onset and offset.

<table>
<thead>
<tr>
<th>Expression</th>
<th>27-point tracking data</th>
<th>51-point tracking data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sessions</td>
<td>Duration (in seconds)</td>
</tr>
<tr>
<td>Anger</td>
<td>65</td>
<td>577</td>
</tr>
<tr>
<td>Anger mimicked</td>
<td>128</td>
<td>1157</td>
</tr>
<tr>
<td>Blink 3 times</td>
<td>31</td>
<td>231</td>
</tr>
<tr>
<td>Disgust</td>
<td>65</td>
<td>574</td>
</tr>
<tr>
<td>Disgust mimicked</td>
<td>128</td>
<td>1153</td>
</tr>
<tr>
<td>Fear</td>
<td>63</td>
<td>555</td>
</tr>
<tr>
<td>Fear mimicked</td>
<td>129</td>
<td>1161</td>
</tr>
<tr>
<td>Happy</td>
<td>68</td>
<td>602</td>
</tr>
<tr>
<td>Happy mimicked</td>
<td>130</td>
<td>1197</td>
</tr>
<tr>
<td>Neutral</td>
<td>68</td>
<td>562</td>
</tr>
<tr>
<td>Neutral mimicked</td>
<td>130</td>
<td>1122</td>
</tr>
<tr>
<td>Open mouth</td>
<td>32</td>
<td>247</td>
</tr>
<tr>
<td>Pout</td>
<td>35</td>
<td>290</td>
</tr>
<tr>
<td>Puff out cheeks</td>
<td>32</td>
<td>271</td>
</tr>
<tr>
<td>Raise eyebrows</td>
<td>30</td>
<td>260</td>
</tr>
<tr>
<td>Sad</td>
<td>66</td>
<td>585</td>
</tr>
<tr>
<td>Sad mimicked</td>
<td>128</td>
<td>1163</td>
</tr>
<tr>
<td>Scrunch up nose</td>
<td>31</td>
<td>251</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1359</strong></td>
<td><strong>11958</strong></td>
</tr>
</tbody>
</table>
Table 6.7: Overview of PD participants dataset by subject. The table shows the number of original frames extracted from video sequences of different PD participants using two facial feature trackers. The Frames (re-sampled) describes the total number of normalised frames.

<table>
<thead>
<tr>
<th>Subject</th>
<th>27-point tracking data</th>
<th>51-point tracking data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC101</td>
<td>194 931 441 62 3173 13180</td>
<td>193 921 414 88 2956 11390</td>
</tr>
<tr>
<td>SC102</td>
<td>496 2486 91230 12250</td>
<td>491 2464 70923 10850</td>
</tr>
<tr>
<td>SC103</td>
<td>502 2411 97261 13150</td>
<td>494 2378 96440 13650</td>
</tr>
<tr>
<td>SC104</td>
<td>539 4479 207338 14950</td>
<td>436 3706 172045 12100</td>
</tr>
<tr>
<td>SC105</td>
<td>194 931 441 62 3173 13180</td>
<td>193 921 414 88 2956 11390</td>
</tr>
<tr>
<td>SC106</td>
<td>522 3970 206479 14150</td>
<td>519 3962 158117 12450</td>
</tr>
<tr>
<td>SC107</td>
<td>519 4437 155018 13100</td>
<td>427 3703 107735 8750</td>
</tr>
</tbody>
</table>

Total: 3173 20662 913499 82600
Table 6.8: Overview of PD dataset with respect to each expression. The table shows the number of original frames extracted from video sequences from PD dataset (by expression) using two facial feature trackers. The *Frames (Re-sampled)* describes the total number of normalised frames.

<table>
<thead>
<tr>
<th>Expression</th>
<th>27-point tracking data</th>
<th>51-point tracking data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sessions</td>
<td>Duration (in seconds)</td>
</tr>
<tr>
<td>Anger</td>
<td>191</td>
<td>1354</td>
</tr>
<tr>
<td>Anger mimicked</td>
<td>170</td>
<td>1275</td>
</tr>
<tr>
<td>Blink 3 times</td>
<td>169</td>
<td>756</td>
</tr>
<tr>
<td>Disgust</td>
<td>191</td>
<td>1175</td>
</tr>
<tr>
<td>Disgust mimicked</td>
<td>170</td>
<td>1291</td>
</tr>
<tr>
<td>Fear</td>
<td>188</td>
<td>1206</td>
</tr>
<tr>
<td>Fear mimicked</td>
<td>170</td>
<td>1244</td>
</tr>
<tr>
<td>Happy</td>
<td>191</td>
<td>1352</td>
</tr>
<tr>
<td>Happy mimicked</td>
<td>170</td>
<td>1289</td>
</tr>
<tr>
<td>Neutral</td>
<td>190</td>
<td>714</td>
</tr>
<tr>
<td>Neutral mimicked</td>
<td>167</td>
<td>752</td>
</tr>
<tr>
<td>Open mouth</td>
<td>169</td>
<td>1063</td>
</tr>
<tr>
<td>Pout</td>
<td>170</td>
<td>1347</td>
</tr>
<tr>
<td>Puff out cheeks</td>
<td>168</td>
<td>1005</td>
</tr>
<tr>
<td>Raise eyebrows</td>
<td>170</td>
<td>1106</td>
</tr>
<tr>
<td>Sad</td>
<td>190</td>
<td>1356</td>
</tr>
<tr>
<td>Sad mimicked</td>
<td>170</td>
<td>1229</td>
</tr>
<tr>
<td>Scrunch up nose</td>
<td>169</td>
<td>1148</td>
</tr>
</tbody>
</table>

Total: 3173 | 20662 | 913499 | 82600 | 2956 | 19063 | 754350 | 72800
Table 6.9: Effect of varying several modes within ±3 standard deviations for several facial expressions.

<table>
<thead>
<tr>
<th>Expressions</th>
<th>Mode index</th>
<th>+3 to -3 SD</th>
<th>Overlapped shapes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disgust</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy mimicked</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy mimicked</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy mimicked</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral mimicked</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fear mimicked</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fear mimicked</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.4. **THE CASE STUDY**

6.4.6 **Observation and Evaluation**

We examined the extent to which measures of facial expressivity can differ between the PD group and controls. As every element of \( c \) describes the degree of facial expressivity using a specific mode in \( Q \) as shown in Table 6.9, our aim was to evaluate how PD and controls score on these modes so the most PD-sensitive modes could be highlighted. We plotted elements of \( c \) against time for each sequence. An example can be seen in Figure 6.5, which shows the variation of the first element of \( c \); the parameter values are shown per frame for both PD (red) and controls (green) when performing a happy expression.

To quantify the most PD-discriminative daily score in a way that can describe PD symptom severity relative to facial muscle movements, we computed several statistics for each element of the parameter \( c \), such as the variance and baseline of the expression curve, as well as the minimum, maximum and maximum distance from the baseline. We also examined differences in the mean magnitude of \( c \) and the expression’s duration from onset to offset. As we examined these parameters, none seemed to evidence clear distinctive boundaries between the two groups. This finding is probably due to the day-to-day variability of PD severity.

Instead, we examined the sample mean distribution of these parameters using the bootstrapping with replacement sampling method with 1000 iterations. Samples were randomly drawn with an equal size as the original data at each iteration. Empirical estimates of the mean and standard deviation were used to approximate their distribution functions. An example of the resulting histograms and their approximated distributions are shown in Figure 6.6.

Furthermore, for each measure, we examined the overlapping distributions between HC and PD groups. Our results suggest that these measures have potential discriminative power. Table 6.10 shows the least overlapping distributions, up to 1%.

The results in Table 6.10 suggest several significant differences between both groups.
However, some of these differences were consistently reported in both 27- and 51-point models. For example, in the happy-mimicked expression, there were three significant measures in each model, all with similar measures. Similarly, measures of peak from baseline and variance were significant for the ‘happy’ expression in both models. Moreover, the duration of the mimicked ‘fear’ expressions was significantly different in both models.

However, the inconsistency in the reported significance of durations was expected since differently shaped models with a varying number of points are not necessarily equivalent, and our annotation methods rely on what defines the major Principal Component (PC). Table 6.9 geometrically demonstrates several movements of these measures.
Figure 6.5: Facial expression variation of PD subjects (red) compared to controls subjects (green) using first mode of Happy model. Units of y-axis correspond to units of standard deviations.
Table 6.10: Least overlapping distributions of examined measures between HC and PD. The table shows measures derived from both 27- and 51-point based features.

<table>
<thead>
<tr>
<th>No. Features</th>
<th>Expression</th>
<th>Measure</th>
<th>Mode index</th>
<th>Overlap ≤ 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anger</td>
<td>Mean Magnitude</td>
<td>-</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Disgust</td>
<td>Mean Magnitude</td>
<td>-</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Sad</td>
<td>Mean Magnitude</td>
<td>-</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Happy mimicked</td>
<td>Variance</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>27 points</td>
<td>Open mouth</td>
<td>Variance</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>Variance</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Happy mimicked</td>
<td>Min</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Disgust mimicked</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Happy mimicked</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Neutral mimicked</td>
<td>Duration</td>
<td>-</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Anger</td>
<td>Duration</td>
<td>-</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Puff out cheeks</td>
<td>Duration</td>
<td>-</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Sad</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Fear mimicked</td>
<td>Duration</td>
<td>-</td>
<td>0.006</td>
</tr>
<tr>
<td>51 points</td>
<td>Happy mimicked</td>
<td>Variance</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Fear</td>
<td>Mean Magnitude</td>
<td>-</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>Variance</td>
<td>0</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Happy mimicked</td>
<td>Min</td>
<td>0</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>Fear mimicked</td>
<td>Max</td>
<td>1</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>Mean Magnitude</td>
<td>-</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Pout</td>
<td>Mean Magnitude</td>
<td>-</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Blink 3 times</td>
<td>Mean Magnitude</td>
<td>-</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>Neutral mimicked</td>
<td>Mean Magnitude</td>
<td>-</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Anger mimicked</td>
<td>Variance</td>
<td>0</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Happy mimicked</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Disgust</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Fear mimicked</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Scrunch up nose</td>
<td>Duration</td>
<td>-</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Fear mimicked</td>
<td>Duration</td>
<td>-</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>Happy</td>
<td>Duration</td>
<td>-</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>Open mouth</td>
<td>Duration</td>
<td>-</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>Neutral mimicked</td>
<td>Duration</td>
<td>-</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>Neutral mimicked</td>
<td>Peak from baseline</td>
<td>0</td>
<td>0.010</td>
</tr>
</tbody>
</table>
6.4. THE CASE STUDY

Figure 6.6: PD subjects (red) and controls (green) sample mean distributions of several measures from different expressions, which shows potential discriminative elements of the parameters $c$ derived from 27-point shapes.
6.4.7 Textural-based Features

We further modelled facial behaviour using several appearance-based features, such as grey intensities, LBP s and Gabor magnitudes, extracted from all frames of the normalised expression sequences. For each frame, we projected each facial image onto two reference frames, where the reference length was matched to 100 and 50 pixels in width. Then, we extracted two rectangular patches around each facial feature (see Figure 6.7) at 21 pixels in width for the LBP feature and 25 pixels in width for the Gabor magnitude.

![Figure 6.7: An example of 27 sampled rectangular patches around different facial feature](image)

For the LBP, we computed patterns in eight neighbourhoods based on gray image intensities. Then, we computed the histograms of all uniform patterns, where each uniform pattern was assigned to a separate bin, while all non-uniform patterns were allocated to a single bin. For the Gabor magnitude, we used a total of 128 Gabor kernels of sinusoidal functions with different wavelengths, orientations and standard deviations of the Gaussian envelope as described in section 4.4.3.1 in Chapter 4. Finally, we combined the geometric feature vector with every appearance feature and created another combined feature vector by concatenating geometrical-based features with the textural-based features of gray intensities, LBP histograms and sequences of...
6.5. CORRELATIONS

magnitudes of Gabor kernels. To reduce dimensionality, we applied a further PCA on all obtained features and computed their vector projections for each sample. Next, we examined the correlation of these parameters with the QoL scores as described in the next section.

6.5 Correlations

The clinical assessment of PD severity is known to be difficult as discussed previously in this chapter, which makes obtaining accurate ground truth data very challenging. Measures of QoL are based on an assessment of individuals’ well-being and are commonly used to evaluate disease progression or the effect of new treatments [241]. Therefore, as we obtained such scores in our case study, we examined the correlation of QoL scores with different features, locally for each subject and globally for all subjects using a regression-based method.

To avoid overfitting, especially with small sample sizes (e.g., with subject-dependant models), we applied Least Absolute Shrinkage and Selection Operator (LASSO) [242]. LASSO is generally used for feature selection and regression regularisation by constraining the sum of coefficients of each predictor to a pre-specified value as shown in Equation 6.1.

\[
\text{minimize}_{\beta_0, \beta} \quad \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2
\]

subject to \( \sum_{j=1}^{p} |\beta_j| \leq t. \) (6.1)

Where \( \beta_0 \) is the intercept and \( \beta_j \) is the coefficient for the \( j^{th} \) predictor \( x_j \). The vector \( x \) is set to contain certain measures, such as the maximum, duration of expression and maximum magnitude, from all the different features. Then, we performed several correlation analyses to identify the best subject-dependant and general regression models.
All measures derived from different features were analysed independently for each expression using the $K$-fold cross-validation procedure with $K=3$ for subject-dependant models and $K=10$ for general models. Since LASSO does not evaluate the significance of the estimate, we used the output feature from LASSO as an input for stepwise linear regression.

Then, we ranked all regression models based on their degree of freedom over the number of predictors to highlight models with higher predictive power and a smaller number of predictors relative to the sample size. Finally, we reported the significant models with $p$-value $\leq 0.05$, adjusted-$R^2 \geq 0.30$. Results of the correlation assessment are presented in Tables 6.11 and 6.12, which shows the top-ranked models for subject-dependant and general models.

Table 6.11: Significant correlations between QoL scores and subject-dependant regression models built from various features extracted around 27 and 51 facial landmarks width ($p$-value $\leq 0.05$).

<table>
<thead>
<tr>
<th>Model points</th>
<th>Subject</th>
<th>Expression</th>
<th>Feature</th>
<th>F-statistic</th>
<th>SE</th>
<th>Adj-$R^2$</th>
<th>MV</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>SC106</td>
<td>Happy mimicked*</td>
<td>N(c)</td>
<td>15.78</td>
<td>1.24</td>
<td>0.35</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>SC102</td>
<td>Disgust mimicked</td>
<td>GM</td>
<td>15.86</td>
<td>0.89</td>
<td>0.5</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>SC103</td>
<td>Disgust</td>
<td>GM</td>
<td>17.16</td>
<td>0.98</td>
<td>0.55</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>SC107</td>
<td>Fear mimicked</td>
<td>GM</td>
<td>63.14</td>
<td>0.33</td>
<td>0.84</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>SC103</td>
<td>Scrunch up nose</td>
<td>I</td>
<td>26.76</td>
<td>1.11</td>
<td>0.7</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>51</td>
<td>SC106</td>
<td>Anger mimicked*</td>
<td>N(c)</td>
<td>11.53</td>
<td>1.21</td>
<td>0.36</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>SC104</td>
<td>Neutral mimicked</td>
<td>LBP</td>
<td>22.83</td>
<td>1.01</td>
<td>0.65</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>SC102</td>
<td>Disgust</td>
<td>F</td>
<td>47.36</td>
<td>0.69</td>
<td>0.79</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

Features: Intensities (I), Local binary patterns (LBP), Normalised $c$ (N(c)), Gabor magnitude (GM) and their fusion (F).

* Models that describe geometrical shape variations.

Since our goal was to identify facial muscle impairments that correlate with QoL scores, we discarded models that do not describe geometrical shape variance (see models annotated with * in Tables 6.11 and 6.12). Moreover, we examined each individual model to assess for violations of multivariate regression assumptions, such as normality and
Table 6.12: Significant correlations between QoL scores and regression models built from various features extracted around 27 and 51 facial landmarks ($p \leq 0.05$).

<table>
<thead>
<tr>
<th>General Models</th>
<th>Models points</th>
<th>Expression</th>
<th>Feature</th>
<th>F-statistic</th>
<th>SE</th>
<th>Adj-$R^2$</th>
<th>MV</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Anger mimicked</td>
<td>LBP</td>
<td>13.11</td>
<td>1.42</td>
<td>0.35</td>
<td>4</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Blink 3 times</td>
<td>LBP</td>
<td>19.28</td>
<td>1.25</td>
<td>0.34</td>
<td>1</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disgust mimicked*</td>
<td>LBP</td>
<td>9.44</td>
<td>1.42</td>
<td>0.32</td>
<td>4</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pout*</td>
<td>GM</td>
<td>10.33</td>
<td>1.35</td>
<td>0.32</td>
<td>4</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sad mimicked</td>
<td>GM</td>
<td>8.13</td>
<td>1.36</td>
<td>0.37</td>
<td>7</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Puff out cheeks</td>
<td>I</td>
<td>10.45</td>
<td>1.35</td>
<td>0.40</td>
<td>5</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scrunch up nose</td>
<td>I</td>
<td>9.42</td>
<td>1.34</td>
<td>0.37</td>
<td>5</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pout</td>
<td>F</td>
<td>8.51</td>
<td>1.35</td>
<td>0.33</td>
<td>5</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>Anger</td>
<td>LBP</td>
<td>8.73</td>
<td>1.43</td>
<td>0.38</td>
<td>7</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anger mimicked*</td>
<td>LBP</td>
<td>11.30</td>
<td>1.39</td>
<td>0.39</td>
<td>6</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fear mimicked</td>
<td>LBP</td>
<td>14.21</td>
<td>1.25</td>
<td>0.45</td>
<td>5</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neutral mimicked</td>
<td>GM</td>
<td>10.11</td>
<td>1.46</td>
<td>0.33</td>
<td>3</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anger mimicked*</td>
<td>I</td>
<td>12.52</td>
<td>1.41</td>
<td>0.38</td>
<td>5</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anger mimicked</td>
<td>F</td>
<td>8.94</td>
<td>1.46</td>
<td>0.33</td>
<td>6</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fear mimicked</td>
<td>F</td>
<td>6.74</td>
<td>1.38</td>
<td>0.33</td>
<td>7</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Open mouth</td>
<td>F</td>
<td>11.49</td>
<td>1.33</td>
<td>0.46</td>
<td>4</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Raise eyebrows</td>
<td>F</td>
<td>7.16</td>
<td>1.38</td>
<td>0.31</td>
<td>7</td>
<td>89</td>
<td></td>
</tr>
</tbody>
</table>

**Included features**: Intensities (I), Local binary patterns (LBP), Normalised c (N(c)), Gabor magnitude (GM) and their fusion (F).

* Models that describe geometrical shape variations.

the equal variance (homoscedasticity) of residuals. For normality, visual tools and statistical tests were used, such as quantile-quantile plot (QQ-plot) and the Shapiro-Wilk test [243]. QQ-plot, under the normality assumption, compares theoretical quantiles of residuals with the estimated values. When the assumption is true, most of these quantiles should follow the diagonal line (i.e., the perfect fit). Conversely, when the residuals are not normal, these quantiles will depart from this line. For the Shapiro-Wilk test, normality is assumed under the null hypothesis. Therefore, if its $p$-value is significant ($p \leq 0.05$), the null hypothesis is rejected, suggesting an abnormal distribution.

We also examined model bias by evaluating influential samples using Cook’s distance [244]. We also validated for homoscedasticity by examining the spread-location plot.
In the subject-specific analysis, most models were filtered out, and only two models based on the N(c) feature for subject SC106 were considered. The happy-mimicked and anger-mimicked models with 27- and 51-points, respectively, showed geometrical shape variation. The happy-mimicked model was based on the variance of the fourth PC, while in the anger-mimicked model, the selected measure was the minimum score of the fourth PC. To understand their geometric properties, we depicted their shape changes within ±3 standard deviations of relevant modes as in Figure 6.14. As shown in the figures, both of these modes describe muscle movements around the mouth. The assessment of multiple linear regression assumptions is presented in Figures 6.8, 6.9 and Table 6.13. In the Residuals vs Fitted plots, all of the clearly non-linear patterns are explained by the model. The rest of the figures also show no violations of the normality assumption. However, the regression line of the happy-mimicked model was found to be sensitive to a single observation as highlighted by Cook’s distance and shown in Figure 6.8.
Figure 6.8: Diagnostic plots for subject-dependent 27-point model of happy mimicked expressions (shape: 27-point, Features: $N(c)$).
Figure 6.9: Diagnostic plots for subject-dependent 51-point model of anger mimicked expressions (shape: 51-point, Features: N(ε)).
Of the 27-point general models, two were correlated with the geometrical feature disgust-mimicked (for LBP feature) and pout (for GM feature). In the disgust-mimicked model variables, the maximum score of the $5^{th}$ PC was associated with shape variations, while in the pout model’s variables, the association was for the minimum of the $7^{th}$ PC. Their effect within $\pm 3$ SD on the shape model can be seen in Figure 6.15.

Of the 51-point models, anger-mimicked was significant in both LBP- and intensities-based models. The geometric effect was associated with the minimum score of the second PC for the LBP model, while for the intensity model, the association was with the vector magnitude. Their assessment of multiple linear regression assumptions are presented in Figures 6.10, 6.11, 6.12 and 6.13, while statistical tests of normality are presented in Table 6.14. Based on these assessments, none of these models violate assumptions of linear regression.
Figure 6.10: Diagnostic plots for general model of disgust mimicked expressions (shape: 27-point, Features: LBP).
Figure 6.11: Diagnostic plots for general model of pout expressions (shape: 27-point, Features: Gabor Magnitudes).
Figure 6.12: Diagnostic plots for general model of anger mimicked expressions (shape: 51-point, Features: LBP).
Figure 6.13: Diagnostic plots for general model of anger mimicked expressions (shape: 51-point, Features: Intensities).
Table 6.13: Normality test of Residuals (subject-dependent models)

<table>
<thead>
<tr>
<th>Model points</th>
<th>Subject</th>
<th>Expression</th>
<th>Feature</th>
<th>W-statistic</th>
<th>p-value</th>
<th>Normal?</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>SC106</td>
<td>Happy mimicked</td>
<td>N(c)</td>
<td>0.94</td>
<td>0.13</td>
<td>YES</td>
</tr>
<tr>
<td>51</td>
<td>SC106</td>
<td>Anger mimicked</td>
<td>N(c)</td>
<td>0.93</td>
<td>0.15</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 6.14: Normality test of Residuals (general models)

<table>
<thead>
<tr>
<th>Model points</th>
<th>Expression</th>
<th>Feature</th>
<th>W-statistic</th>
<th>p-value</th>
<th>Normal?</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Disgust mimicked</td>
<td>LBP</td>
<td>0.98</td>
<td>0.33</td>
<td>YES</td>
</tr>
<tr>
<td></td>
<td>Pout</td>
<td>GM</td>
<td>0.98</td>
<td>0.16</td>
<td>YES</td>
</tr>
<tr>
<td>51</td>
<td>Anger mimicked</td>
<td>LBP</td>
<td>0.98</td>
<td>0.18</td>
<td>YES</td>
</tr>
<tr>
<td></td>
<td>Anger mimicked</td>
<td>I</td>
<td>0.99</td>
<td>0.80</td>
<td>YES</td>
</tr>
</tbody>
</table>

Figure 6.14: Significant subject-dependent models with correlated facial movements. m: represent mode index.

(a) Happy mimicked (N(c), m=3) (b) Anger mimicked (N(c), m=3)

Figure 6.15: Significant general models with correlated facial movements. m: represent mode index.

(a) Anger mimicked (LBP, m=1) (b) Disgust mimicked (LBP, m=4) (c) Pout (GM, m=6)

6.6 Discussion and Conclusion

In this chapter, we presented a review of the literature on facial expression analysis in PD with a focus on facial expressivity assessment. Many studies have analysed PD
6.6. DISCUSSION AND CONCLUSION

cognitive impairment, while few attempts have been made to analyse facial expressivity in PD. Although a few studies have implemented fully automatic methods, none of these methods were evaluated in naturalistic settings.

We used a novel unsupervised-based method to learn about expression manifolds. The video recordings of PD participants were taken under unconstrained conditions with no special settings on recording devices. This led to technical challenges related to the production of multiple video formats with varying frame rates. These problems were resolved by converting the video files to a common format (MPEG) and frame rate. Consequently, the original quality of these videos was not retained in our software, limiting its full potential.

The assessment of current tracked facial features showed that utilising RFCLM and AAM models to track facial features in naturalistic settings was sufficient for the purpose of automatic facial behaviour analysis with most subjects. However, the tracked data for subject SC107 were associated with a large loss of frames. This was due to extremely unusual severe illuminations that rendered most facial features invisible even to the naked eye. The frames were discarded by the facial tracker due to their failure to satisfy the quality-of-fit threshold.

From expression manifolds, we explored several statistical measures to examine differences between controls and PD subjects. Our results highlight several significant differences, such as expression duration, intensities (i.e., peaks from the global mean and the subject’s baseline) and variances. In terms of the number of significant measures associated per expression, happy and happy-mimicked were shown to have the most associations with the highlighted parameters and were consistent in both the 27- and 51-point models.

Moreover, the duration of neutral-mimicked expressions was significantly different between the two groups in both models ($p \leq 0.01$). One plausible reason for this is that the patient’s decoding of neutral faces is more disturbed in PD compared to controls as
suggested in previous findings [245]. Thus, misperceiving emotional signals can trigger false emotional responses [246].

Since our goal was to highlight correlated parameters with QoL scores and because the number of predictors was greater than the number of samples \((p \gg n)\), we used LASSO for the initial feature selection. Then, we used stepwise linear regression to examine the correlation between these parameters.

Linear regression was performed twice for each expression, independently based on subject-specific and overall data. In our results, several significant models are highlighted; however, some models are associated with very low-rank PC, which is essentially noise. Therefore, we filtered models that were not associated with large geometric variations.

For 27-point subject-specific models, only one model (happy-mimicked) was considered. Similarly, for the 51-point models, a single model (anger-mimicked) was considered. We notice that both models belong to the same subject (SC106), which has the largest number of samples compared to other subjects. However, diagnostic plots of residuals have shown that the happy-mimicked model may fail to generalise due to an influential case, while anger-mimicked as well as the 51-point general models did not show this problem.

In the 27-point models, disgust-mimicked and pout expressions were significantly correlated with QoL. The 51-point mimicked expressions of anger are found more often correlate with PD QoL than self-posed expressions when analysed using a 51-point model. To understand the geometric interpretations, we showed their overlapping shapes when varying relative PC between \(\pm 3\) SD as shown in Figures 6.14 and 6.15.

Discriminative measures of facial expressions do not necessarily correlate with QoL scores. The best test for evaluating QoL, as highlighted in our case study, is by mimicking the emotion of anger (in a 51-point analysis) and disgust (in a 27-point analysis). Self-posed pout can also be used with 27-point models to assess QoL.
Chapter 7

Discussions and Future Work

The main objective of this research was to study the dynamics of facial behaviour in recorded video sequences under naturalistic settings to aid clinical assessment of facial expressivity. Feeding the model with reliable data is crucial for the analysis step. Therefore, in this work we demonstrate how we optimise several parameters of two facial tracking techniques generic (RFCLM) and subject-specific (AAM) by extensive experiments in a multi-stage manner. We evaluate the final optimised generic tracker on several publicly available datasets including the challenging dataset of The-300-Faces-in-the-wild challenge (300W)[247]. Even though we did not achieve the best result on 300W[58] dataset, our baseline of initial error is higher and we still achieve a comparative results with recent state-of-the-art methods.

Further, we developed a novel method to model facial expressivity using several facial features (geometrical, textural and their fusion). To validate our model, we conducted a case study to assess Parkinson’s Disease (PD) effects on facial expressivity. Since obtaining a ground truth assessment of PD is still challenging even for a neurologist due to symptoms’ similarity with other diseases, the best ground truth score we could obtain was the assessment of daily quality-of-life score derived from participants’ self-ratings on a questionnaire, which was designed by a specialist.
Another contribution of this project is based on dataset labelling with meta information. This includes annotating each video segment with an approximate estimation of the expression onset and offset, which is later refined using an automatic refinement step in our framework. This enabled an efficient processing of further feature extraction and analysis using only a few samples that represent facial dynamics from-and-to the baseline (neutral state).

The number of model parameters derived from different features was larger than the number of observations for each subject and expression. Therefore, to avoid overfitting in the regression analysis, we applied two methods. First we used LASSO as a feature selector which constrains the number of allowed predictors by enforcing a regularisation penalty on the absolute sum of coefficients. Then all selected parameters are input to stepwise linear regression to assess their significance. We also validated our regression model by examining the distribution of the plot of residuals against fitted values. The results suggest there were no obvious non-linear patterns to be further explained by the model.

Our findings suggest that our method can predict significant model parameters which can explain the variation of the quality-of-life score in PD subjects. We validate our resulting models using several methods such as LASSO to reduce overfitting and stepwise linear regression to assess the significance of those models. We also applied visual and statistical tests of normality to validate the derived correlations. We also found that the most impaired expression does not necessarily best describe the quality-of-fit scores of PD subjects.

In this thesis, we optimised two types of facial tracking techniques: generic RFCLM and person-specific AAM. We evaluated RFCLM trackers on several in-the-wild public datasets. We showed RFCLM performance in comparison with several state-of-the-art methods. The RFCLM did not outperform the current state-of-the-art. However, unlike CNN-based methods, our approach is more efficient and less complex. The AAM models were used for subjective tracking we did not aim to establish a generic
AAM trackers since RFCLM were sufficient for the behavioural analysis task for most subjects.

We assessed tracking performance in chapter 6. The 51-point tracker performed less well than the 27-point, however, this was due to high constraint imposed on quality-of-fit value. This high constraint allowed us to track frames with less errors. This has improved the learning of expression subspace as shown in Table 6.9, where the model describes more expressive shapes.

In the gaze chapter, we compared the use of three different features: SSM, (SSM + eyes-corners distances) and integral images of Haar-like features. Our preliminary results show that combined SSM and eyes-corners distance can improve gaze prediction performance.

7.1 Challenges and Future work

7.1.1 Challenges

One of the main challenges faced in this project was the degraded video quality due to two main reasons. First, sometimes the automatic video recording that is generated by the machine native environment, failed to include important video meta information such as the frame rate and the duration of the video, which is critical for the frame extraction step. To solve this problem we were forced to convert the video to a different format at the cost of degrading video quality especially when the original resolution was not high. Another effect that reduces the video quality is recording in a very low light condition. Some participants performed their recording indoor at night time, in a low light environment. This affects the image sensor of the camera, which creates a grain effect on the image.

Due to project time constraints and changes in research directions we could not extend our study to include the gaze model in behaviour analysis. One major time-consuming
task was the labelling of each expression onset and offset. The PD dataset alone contains a set of 7020 different expressions. Even though we took extra care in labelling these data sometimes we were forced to repeat these annotations due to a slight time shift of converted format.

7.1.2 Future work

- We evaluate our facial behaviour model to assess the severity of Parkinson’s disease represented in the quality-of-life score. Future work could be dedicated to examine psychotic disorders which are associated with facial expression impairment including schizophrenia and Bipolar disease.
- Since the best estimate of PD severity used in this project is the quality-of-life score, this measure may not well represent the effect of PD on facial expressivity. Therefore, further measures could be examined to answer this question.
- We had access to only seven people with PD, limiting our ability to draw conclusions about the general population with the disease. Future work would involve collecting data from a much larger cohort, and over a longer period.
- Plans for a study of people with Schizophrenia were delayed, so the data was not available for study in this project. When this is collected it could be analysed using the same approach as we used here.
- In facial feature and gaze tracking, there is a significant shift toward deep learning and CNN-based methods. This can be attributed to their success in many computer vision problems. Our review showed that most state-of-the-art methods are CNN-based. Future work could be carried out to evaluate PD facial masking using these approaches.
Bibliography


[82] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De La Torre. Detecting depression from facial actions and


[85] C. Bell. The anatomy and philosophy of expression as connected with the fine arts. George Bell & Sons, 1904.


(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 656–667, 2008.


[168] A. Nakazawa and C. Nitschke. Point of Gaze Estimation through Corneal Surface


[201] K. R. Chaudhuri, D. G. Healy, and A. H. Schapira. Non-motor symptoms of


