COMPUTATIONAL MODELLING TO TRACK HUMAN EMOTION TRAJECTORIES THROUGH TIME

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

at School of Engineering and Advanced Technology,

Massey University, Palmerston North,

New Zealand.

AYESHA HAKIM

2013

Acknowledgements

I owe the most sincere gratitude to my supervisors, Prof. Hans Guesgen and Prof. Stephen Marsland, for the continuous support during my Ph.D. study and research. Their patience, kind guidance, friendship, and immense knowledge helped me not only in my research and thesis writing, but also to deal with the daily-life challenges. Of course, it is not easy to do a PhD with a young family! I couldn't have asked for more supportive advisors and talented teachers.

Thanks to all the members of Massey University Smart Environment (MUSE) group for their encouragement, insightful comments, and valuable feedback on my research and presentations. Thanks to Michele Wagner for all the administrative support throughout my degree.

I would like to thank my office-mate and friend, Le Thu, for keeping my spirits high and giving useful advise to cope with personal and study-life problems. Many thanks to An, Linda, and Amissa for providing me with an ear to listen after hard comments from the supervisors. We did a lot of back-biting!

Heaps of acknowledgement go to Higher Education Commission, Pakistan for funding my research. Thanks to Universities New Zealand for granting me the Claude McCarthy Fellowship to participate in the conference. Thanks to Dilantha for kind guidance related to account issues, and Natalia for her friendly attitude and kind support for university issues. Many thanks to the authors of IEMOCAP dataset for

providing me access to the continuous unsegmented data. A special thanks to Prof. Carlos Busso for helping me understand and use this dataset throughout the research.

Last but not the least, I would like to thank my parents Abdul Hakim Khanzada and Nasira Hakim and sisters Sahar, Mahrukh, Rabia, who encouraged me to go for Doctoral studies and gave me full support mentally and emotionally throughout my PhD, and my life. Many thanks to my husband Tariq and son Ali, who gave me happy moments after a tiring day at Massey. I love and respect all of you from the depth of my heart.

Abstract

There has been a lot of research into the field of affective computing over the past three decades. In the context of this thesis, affective computing is the computing that relates to emotion recognition, representation, and analysis. Much of the past work has focused on the basic emotions. However, most human emotions are not pure examples of one basic emotion, but a mixture of them, known as complex emotions. Emotions are dynamic, they change continuously over time. This thesis focuses on computational modelling to recognise, represent, and analyse continuous spontaneous emotions through time.

Emotions are internal, and hence impossible to see directly. However, there are some external presentations of emotions enabling computational tools to be used to identify them. This thesis focuses on the use of facial points as a measure of underlying emotions. The main focus is the development of computational models to track the patterns of facial changes in order to analyse the paths followed by emotions over time.

While there has been lots of work on shape models to classify facial expressions into discrete basic emotion categories, they are generally based on the analysis of the full face. However, the research shows that some expressions are better recognized by muscle activity in the upper half of the face, while others use muscles primarily from the lower half of the face. This thesis introduces a joint face model based on

shape models of full, upper, and lower parts of the face separately that significantly improves the accuracy.

The set of shape models gives a degree of match to each basic emotion. Using this information, this thesis addresses the problem of complex emotion recognition by developing a mixture model that combines each basic emotion in an appropriate amount. The proposed model represents emotions in the activation-evaluation space, which is the most widely-used representation of emotions in psychological studies. It represents emotions on the basis of their polarity and similarity to each other. This thesis uses a mixture of von Mises distributions for emotion recognition, which is an approximation to the normal distribution for circular data and is the most common model for describing directional data. The results show that the proposed mixture model fits the data well.

Emotions vary continuously with regard to intensity, duration, persistence with time, and other attributes. In addition, their appearance on the face varies, and the transition in facial expressions is based on both the change in emotion and physiological constraints. This thesis examines the trajectories between emotions in activation-evaluation space and shows that these trajectories are smooth and follow 'common' paths between different emotions. In the past, very few efforts have been made on the analysis of continuous emotion dynamics. The findings presented in this thesis can be used and extended in several directions to improve the emotion recognition as well as emotion synthesis.

For my papa and mama

Contents

Acknowledgements **v**

Abstract **vii**

1 Introduction **13**

1.1 Motivation 13

1.2 The Problem 15

1.3 Scope of the Study 17

1.4 Aims, objectives, and hypotheses 18

1.5 Experimental Methodology 20

1.6 Main Research Contributions of Thesis 21

1.7 Overview of Thesis 23

2 Emotions from the Perspective of Psychology **27**

2.1 Introduction 27

2.2 Understanding Emotions 28

2.3 Basic and Complex/Mixed Emotions 29

2.4 Emotions and Facial Expressions 31

2.5 The Space of Emotions 32

2.5.1 Circumplex models of emotions 33

2.5.2 Activation-evaluation space 35
2.6 Categorical and Attribute-based Descriptions 38
2.7 Summary 39

3 Literature Review **41**
3.1 Introduction 41
3.2 Direct Classification Approaches 42
3.2.1 Template-based techniques 42
3.2.2 FACS-based techniques 45
3.2.3 Rule-based techniques 47
3.3 Mapping to Emotion Space Approaches 48
3.3.1 Quantised approaches 48
3.3.2 Continuous approaches 51
3.4 Analysis of Emotion Dynamics 54
3.5 Summary 55

4 Datasets for Emotion Recognition and Analysis **57**
4.1 Introduction 57
4.2 Criteria for Selecting a Dataset 58
4.3 Review of Datasets 59
4.4 Data Annotation Tools 68
4.5 The Selected Dataset 71
4.6 Description of IEMOCAP 73
4.7 Data Preprocessing 76
4.8 Summary 80

5 Basic Emotion Recognition **81**
5.1 Introduction 81

5.2 Shape Models . . . 81
5.2.1 Using the upper and lower face separately . . . 85
5.2.2 Classification . . . 88
5.3 The Effects of Each Principal Component . . . 89
5.3.1 Effect of each full, upper, and lower face PC . . . 90
5.4 Performance Comparison . . . 94
5.4.1 A rule-based emotion classifier . . . 94
5.4.2 Support Vector Machines . . . 98
5.5 Results . . . 98
5.5.1 Robustness to mislabelled data . . . 105
5.6 Discussion . . . 105

6 Statistical Modelling of Complex Emotions using Mixtures of von Mises Distributions 109

6.1 Introduction . . . 109
6.2 Analysis of Circular Data . . . 110
6.2.1 Circular distribution . . . 111
6.2.2 Statistical approaches to modelling circular data . . . 112
6.2.3 Selection of an appropriate circular distribution . . . 114
6.3 Description of Our Method . . . 115
6.3.1 Mapping emotions to the activation-evaluation space . . . 116
6.3.2 von Mises Mixture Model . . . 117
6.3.3 Estimating the Parameters of the Mixture Model . . . 119
6.3.3.1 Estimating the concentration parameter . . . 119
6.3.3.2 Estimating the weights . . . 120
6.4 Experimental Results . . . 120
6.5 Summary . . . 128

7 Computational Analysis of Emotion Dynamics **131**

7.1 Introduction 131

7.2 Hypotheses 132

7.3 Evaluation 133

7.4 Validation 144

7.5 Conclusion 144

8 Conclusions **149**

8.1 Research Overview and Significant Findings of the Thesis 149

8.2 Future Research 157

8.2.1 Using full facial images instead of marker locations 157

8.2.2 Emotion recognition using multiple modalities 158

8.2.3 Adding contextual information 159

8.2.4 Observing the time-offset of emotion activation and recovery . 160

8.2.5 Life-long learning 161

8.3 Conclusions 163

A List of Publications **165**

B Emotion Words from Whissell and Plutchik **167**

C The effects of varying the second principal component on the female mean upper face **171**

D Glossary **175**

References **177**

List of Tables

2.1 List of basic emotions by various theorists. Presented by Ortony and Turner [163]. 30

2.2 Emotion classes in the activation-evaluation space [221] 37

4.1 A review of existing visual (V) and audio-visual (AV) datasets for human emotion recognition and analysis with respect to the presented criteria. 67

5.1 The effect of each PC on the mean face. For explanation, see the text. 90

5.2 Rules for classifying emotions based on the direction and magnitude of the selected full face principal components and the corresponding vocabulary of FAPs associated with the expression of each basic emotion. The vocabulary of FAPs is taken from [181]. The description of each of the mentioned FAP is listed in Table 5.3. The relationship between PCs and the corresponding FAPs is discussed in the text. 95

5.3 Description of each MPEG-4 FAPS mentioned in Table 5.2. 96

5.4 Mean accuracy and standard deviation of the joint face model, full, upper, lower face models, SVM Classifier, and the rule-based Classifier on both 4 and 6 emotion classes. 101

6.1 Angular values from Whissell's study and those estimated by the models. 122

7.1 Coefficient of determination (R^2) of linear, quadratic, and cubic polynomial regression models fitted to the fifteen symmetric paths of emotion transitions into the activation-evaluation space. The emotion categories are denoted as *Neu* for Neutral, *Ang* for Anger, *Fru* for Frustration, *Hap* for Happiness, *Exc* for Excitement, and *Sad* for Sadness 141

B.1 Emotion Words from Whissell and Plutchik. List taken from [44]. The first two numerical values represent valence and activation of each emotion word that was found in the study by Whissell [221], and the fourth column represents the corresponding angular location in the activation-evaluation space that was found in the study by Plutchik [174]. The values of valence range from 1 (negative extreme) to 7 (positive extreme), and the values of activation range from 1 (very passive) to 7 (very active). 170

C.1 The numerical values demonstrating the direction of 3D movement of 17 marker points effected by varying the second PC between $\pm 3\sigma$ on the female mean upper face. μ denotes the mean face and σ denotes the standard deviation. Considering the mean face as a reference, the values identify the upward movement of forehead and eyebrows marker points by varying PC2 from -1σ to -3σ and the downward movement of forehead and eyebrows marker points by varying PC2 from $+1\sigma$ to $+3\sigma$. 173

List of Figures

1.1 A simple illustration of four temporal phases (neutral, onset, apex, and offset back to neutral) of a facial expression of emotion. 16

2.1 A circumplex structure of emotions obtained on the basis of similarity measures. Picture taken from [174]. 34

2.2 A bipolar circumplex of emotions, commonly known as activation-evaluation space. Picture taken from [234]. 37

4.1 Example of FeelTrace display during a tracking session. Cursor colour changes from red/orange at the left hand end of the arc, to yellow beside the active/passive axis, to bright green on the negative/positive axis, to blue-green at the right hand end of the arc [43]. 69

4.2 Iconic representation of valence (top row), activation (middle row), and dominance (bottom row) by Self Assessment Manikins [127]. 70

4.3 Marker Layout used in recording for IEMOCAP dataset [25]. 74

4.4 The ANVIL annotation tool used for emotion evaluation of the utterances based on categorical and continuous attributes [25]. 75

4.5 The distribution of data for each emotion category, *Neu*: Neutral, *Dis*: Disgust, *Hap*: Happiness, *Sur*: Surpsie, *Sad*: Sadness, *Ang*: Anger, *Fea*: Fear, *Fru*: Frustration, *Exc*: Excitement, *oth*: Other [25]. 77

4.6 The 28 marker points in 3D used for emotion recognition and analysis. 78

4.7 An example of data layout in a continuous conversation between two actors (male and female). The term 'overlap' refers to the situation where both actors are talking at the same time. 80

5.1 The scree graph for the full face data plotting percentage variance covered by the first ten principal components. 85

5.2 The effect of varying the first PC (PC1) for the full face of female actor. μ denotes the mean face and σ denotes the standard deviation. 86

5.3 The 17 marker points in 3D on the mean upper face of the female actor. 87

5.4 The 11 marker points in 3D on the mean lower face of the female actor. 87

5.5 The scree graphs plotting the percentage variance covered by first ten principal components for (a) the upper face data, and (b) the lower face data. 87

5.6 The proposed joint face model. 89

5.7 The effects of varying the second principal component (PC2) on the female full face model. μ denotes the mean face and σ denotes the standard deviation. 91

5.8 The effects of varying the second principal component (PC2) on the female upper face. μ denotes the mean face and σ denotes the standard deviation. The numerical values demonstrating the direction of variation of marker points are listed in Table C.1 in Appendix C. . . 92

5.9 The effects of varying the second principal component (PC2) on the female lower face. μ denotes the mean face and σ denotes the standard deviation. 93

5.10 Positions of images in the space of the first three principal components of some of the transformed emotion clusters of full face data in 4D space. All the figures are in 3D, but to get a clear view the viewpoint has been rotated by setting the azimuth and elevation to some suitable value. The figure needs coloured print to differentiate between different clusters. 99

5.11 The comparison of the mean accuracy of the four shape models, the rule-based classifier, and the SVM-based classifier on the female testset. Lines mark one standard deviation. 102

5.12 The comparison of the mean accuracy of the four shape models, the rule-based classifier, and the SVM-based classifier on the male testset. Lines mark one standard deviation. 103

5.13 The robustness of the joint face model and the SVM-based classifier to mislabelled training data. 104

5.14 Some examples of emotion blends. 106

6.1 Illustration of the resultant vector direction and length (in grey) obtained by vector addition of (a) $60°, 180°, 300°$, (b) $120°, 180°, 240°$, and (c) $150°, 180°, 210°$. All the angles start from the horizontal zero-direction. The light grey circle is the unit circle. In (a), we cannot see the resultant vector of length zero, it resides at the centre of the circle. This picture is taken from [19]. 111

6.2 The probability density plots for three different sets of datapoints modelled using wrapped normal distributions with selected parameters μ and σ^2. The red circles show the density estimate based on the observed data. Picture taken from [87]. 113

6.3 The probability density plot for 100 datapoints randomly drawn from a wrapped Cauchy distribution. The red circle shows the density estimate based on the observed data. Picture taken from [87]. 114

6.4 The effect on the density of von Mises distribution for (a) fixed mean direction μ and varying concentration parameter K and (b) fixed concentration parameter K and varying mean direction μ. 118

6.5 The position of each basic emotion (based on the training set) in the activation-evaluation space. 121

6.6 (a) von Mises Probability Distributions in the mixture model with unit weight, (b) von Mises Probability Distributions characterising one frame of an utterance labelled as [angry, angry, frustrated] by three human experts. 123

6.7 The test of fit for (a) the mean ground truth *directions* and those estimated by the mixture model (b) the mean ground truth *intensities* and those estimated by the mixture model, for each of the seven conversations in the test set. Lines mark one standard deviation. . . . 124

6.8 Illustration of the Kuiper's statistic. Blue line is model estimated CDF of (a) anger, (b) frustration, (c) happiness, (d) excitement, and (e) sadness, red line is an empirical CDF of the first conversation in the testing set, and the green line is the Kuiper's statistic. 126

6.9 The mapping of continuous emotions during an utterance on the activation-evaluation space, along with the mean ground truth direction and intensity and those estimated by the model. The movement of emotions through time is represented by changing colour spectrum from dark/red (start) to light/yellow (end). 127

7.1 The mapping of continuous emotions during a conversation (Ses01F_impro01) into the activation-evaluation space. The movement of emotions through time is represented by changing colour spectrum from dark/red (start) to light/yellow (end). 132

7.2 Test of smoothness of emotion paths corresponding to each conversation in the testing set. For explanation, see the text. 134

7.3 A linear regression model fitted to R/S analysis for all emotion trajectories in the testing set. 136

7.4 The size of the 'change' between two consecutive emotion points in the activation-evaluation space, for all conversations in the testing set. The maximum possible motion is 2, since the circle is radius 1. 137

7.5 The false positives appear during transition from neutral to excited. The movement of emotions through time is represented by changing colour spectrum from dark/red (start) to light/yellow (end). 138

7.6 Transitions between negatively correlated emotions tend to pass through the neutral state, while transitions between positively correlated emotions do not. 139

7.7 Ses01F_impro01, continuous frames during emotion transition from anger to happiness. The temporal relationship between intensity change and angle change can be seen. For explanation, see the text. 143

7.8 Outputs of the emotion trajectories for the four different testing sets in the activation-evaluation space. 145

7.9 A synchronised frame-by-frame comparison of videos from the dataset and the corresponding mapped emotions into the activation-evaluation space. The snapshot shows the *laughing* expression during a happy conversation mapped as *excitement* of high intensity on the space. The female actor was being recorded. 146

7.10 The sarcastic expression lead to an abnormal transition towards happiness during an angry conversation. The female actor was being recorded. 147

8.1 Ses01F_impro01, continuous frames (6500-7100) during emotion transitions. Time-offset between intensity change and angle change. . . . 161

8.2 The temporal structure of emotional life, figure taken from [44]. . . . 162

Chapter 1

Introduction

This first chapter provides an overview of the research background, outlines the problems being addressed, and identifies the scope of this thesis. It also describes the aims and objectives, along with some hypotheses. The chapter ends with a brief description of the experimental methodology followed in this thesis, together with an overview of the rest of the chapters.

1.1 Motivation

Humans tailor their interpersonal relationships by recognising emotions. This helps them cope with specific situations, such as realising when somebody else is annoyed. Research findings signify the importance of emotions in learning, decision making, and rational thinking [171]. Based on this, the ability to recognise and express emotions are essential for natural communications between humans.

Nowadays, most of us spend more time interacting with computers than with other humans [171]. Computers control a significant part of our lives and we expect them to be reliable, predictable, and intelligent with rational judgment. We want them to be able to understand what we feel and to adapt accordingly. In order

to communicate intelligently with us, computers will need the ability to recognise, express, and respond to our emotions.

The term 'Affective Computing', coined by Picard, is defined as, 'computing that relates to, arises from, or deliberately influences emotions' [171]. In a short time, it has been a well-accepted interdisciplinary field of research combining psychology, linguistics, anthropology, sociology, behavioural studies, and human-computer interaction, with a common aim of developing technology that serves the human in more sentient and sensible ways [170]. It should be noted that throughout this thesis, terms such as 'affect' and 'emotion' are used interchangeably. Also, affective computing and emotive computing refers to the same field of research.

Having computers recognise and understand emotions has many benefits. These applications include human-robot interaction [14,15], emotional conversational agents [20, 225], sensitive talking heads [105], and sensitive artificial listeners [198]. Other benefits include the health monitoring of patients, e.g., stress/pain monitoring [91, 152, 211], observing and assisting psychiatric patients such as those suffering from schizophrenia, autism, insomnia, and depression [55]. Monitoring drowsiness or state of alertness, panic and disress level of automobile drivers may help prevent serious accidents [79]. In the field of education, emotion-sensitive tutors may interactively adjust the speed and content of tutorial by detecting the sign of boredom or confusion on student's faces [88]. Detecting the student's level of understanding may also assist distance learning/tele-teaching [182]. Other applications include automatic retreival of emotional videos in multimedia [206], personalised gaming in the field of entertainment [183], and software quality assurance by monitoring the signs of frustration or disturbance of users while interacting with the software [30].

1.2 The Problem

Most of the research related to automatic emotion recognition focuses on giving computers the ability to recognise discrete basic emotions using static facial images of posed facial expressions (see Section 3.2). However, research in psychology and related fields has shown that our real-life emotions are not pure examples of one basic emotion, but a mixture of them, known as complex emotions. Basic emotions are combined with different 'weights' to give rise to several *blends* of emotions. Also, the posed emotions are quite different from natural real-life emotions in duration, intensity, and facial muscle actions. For example, spontaneous smiles are of smaller amplitude and have a more consistent relation between amplitude and duration than posed smiles [33,78,215]. Similarly, it has been shown that spontaneous muscle actions have different temporal and morphological characteristics than the posed ones [215]. Consequently, methods of automatic human emotion recognition and analysis trained on deliberate and often exaggerated emotions usually fail to handle complex spontaneous emotional behaviour [233].

The research related to the study of 'discrete emotion dynamics' focuses mainly on the detection of four temporal segments: neutral, where there is no sign of activation of any facial expression; the onset of a facial expression, when the muscular contraction begins and increases in intensity; the apex, which is the peak where the intensity reaches a stable level; and the offset, which is the relaxation of the muscular action back to the neutral state (see Section 3.4). Fig. 1.1 shows the neutral-onset-apex-offset-neutral phases of a facial expression.

However, emotions are not discrete: they continuously change over time due to their natural progression, external stimuli, and the way the face works. Naturally, emotions fade in their intensity with time [185]: the intense anger that might have been accompanied by betrayal by a close friend might provoke a milder response when

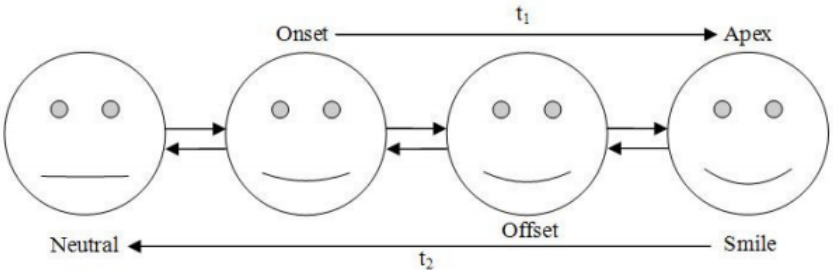


Figure 1.1: A simple illustration of four temporal phases (neutral, onset, apex, and offset back to neutral) of a facial expression of emotion.

recalled after weeks or months. Similarly, the intense feeling of joy accompanied by winning a championship might provoke a milder sense of happiness when looking at the event's photos weeks or months later. Also, research shows that a noticeable change in emotions is brought about by an external stimulus, e.g., the behaviours of others, or a change in the current situation, or internal stimuli such as thoughts or memories [193].

In addition, emotional *expressions* change based on the mechanical properties of facial skin. The facial dermal tissues comprise collagen (72%) and elastin (4%) fibres which help resist deformation of tissues. Therefore, the facial tissues effected by the active muscle activity caused by emotion expressions need to relax before stretching to another form (i.e., expressing another emotion) [207]. Moreover, emotions are related to each other in a systematic manner which guides the way emotions go through transition from one state to another. For example, the transition from anger to frustration is more common than the transition from anger to happiness.

So far, none of the computational studies have focused on the temporal analysis of 'continuous emotions dynamics' in order to understand the relationships between different emotions and the paths followed by emotions while moving from one state to another (i.e., emotion trajectories).

This thesis addresses these problems by taking into account continuous spontaneous expressions of emotions. It focuses on the study of complex emotions as the weighted set of basic emotions, considering their relationships with each other. In addition, this thesis presents ways to study the temporal dynamics of continuous emotions in order to get insight about the emotion trajectories and the corresponding intensity variations over time.

1.3 Scope of the Study

In this thesis, we restrict our scope of research to the computational modelling of facial points and their mapping to some appropriate space and study the trajectories followed by them in this space while a person experiences a set of emotional states through time.

Emotions are internal, and hence impossible to see directly. However, there are generally accepted physical correlates, principally facial expression and tone of voice, although words, hand writing, physiological variations (e.g., change in body temperature, skin conductivity, blood flow, etc.), and brain activations are also correlated with emotions [144]. In 1968, Mehrabian [148] indicated that the verbal part (i.e., spoken words) of a message contributes to 7% of the overall message affect, the vocal part (i.e., voice information) contributes to 38%, while facial expressions contribute to 55% of the affect of an emotional communication.

Among all other presentations of emotions, facial expressions are the most visible and easily available. It is therefore sensible to observe them to assess the underlying emotional state. In this thesis, we further simplify the problem by using the facial points which can capture the facial movements while expressing emotions. These points can be based on the Feature Points (FPs) defined in the MPEG-4 Facial

Animation (FA) standards [165] and are considered to be the best way of representing a human face.

The reason for this simplification is to avoid pre-processing steps such as the extraction and tracking of facial features useful for emotion classification. The choice of facial points instead of full faces enables us to focus on the computational modelling of emotions, which is the main objective of this thesis.

1.4 Aims, objectives, and hypotheses

The aim of this research is to develop a computational system that, when given a set of facial points, can map them to some appropriate emotion space, and analyse the trajectories followed by them while a person experiences a set of emotional states through time. The facial points are assumed to be correlated with the underlying emotions. This thesis focuses on the continuous spontaneous expressions of emotions rather than their discrete posed displays.

In order to achieve this goal, a set of objectives has been formulated. They will be revisited in the concluding chapter. There are three main objectives. The first focuses on mapping the facial points to the weighted set of discrete basic emotions. Since real-life emotions are not discrete, the second objective focuses on complex emotion recognition by representing basic emotions in some appropriate space. Emotions are also not static, so the third objective considers the continuous change in facial points as a time series in order to analyse emotion dynamics through time. Further description is given below:

1. *Mapping facial points to weighted basic emotions*: The first objective is to classify the facial points with regard to the weighted basic emotions. This includes:

    - Choose a set of basic emotions.

- Model the facial points to identify basic emotions.
- Evaluate the results by comparing with the ground truth data of categorical labels assigned by human evaluators. Also, evaluate the performance accuracy and reliability of results by comparing with those obtained by state-of-the-art methods of emotion recognition.
- Describe each frame/set of facial points as a weighted combination of the basic emotions.

2. *Represent emotions in some space*: The second objective is to represent basic emotions on the basis of their associated weights in an appropriate space. It will enable the computers to represent each frame/set of facial points as a 'blend' of basic emotions. This includes:

   - Find a suitable space for representing emotions, with respect to the psychological literature.
   - Develop a method for the mapping of basic emotions into that space.
   - Develop a method to calculate the weighted average of basic emotions in order to represent each frame/set of facial points as their 'blend' in that space.
   - Evaluate the results by comparing with the ground truth data of labels in terms of emotion attributes assigned by human evaluators.

3. *Consider the representation of emotions in the space as a time series*: The third objective is to consider the representation of emotions in the space as a time series in order to analyse the paths followed by emotions while moving from one state to another. A time series is a sequence of values at successive time points

following a non-random order. We propose the following hypotheses about the emotion paths representation in the activation-evaluation space:

- The paths among emotions form 'smooth' trajectories in the space.
- If the end-point emotions are not positively correlated, then the path goes through the neutral state.
- If the end-point emotions are positively correlated, then the path does not go through the neutral state.

For more details about these hypotheses, see Section 7.2. The third objective of this thesis includes:

- Test the proposed hypotheses to study the dynamics of emotions in a time series.
- Evaluate the results by cross-validation using the ground truth data of discrete categorical labels as well as in terms of continuous emotion attributes.

In all the objectives stated above, we test our methods on a spontaneous emotion dataset called the Interactive Emotional Motion Capture (IEMOCAP) dataset (see Section 4.6) to investigate the ability of these methods to achieve these objectives. The contents of Chapters 5, 6 and 7 reflect each of the research objectives described above.

1.5 Experimental Methodology

For the first and second objectives stated in the previous section, we plan to develop models to predict results, and test the accuracy and reliability of those results by

comparing with the ground truth data labelled by three human evaluators (which is given in the IEMOCAP dataset).

In this dataset, a pair of actors were recorded using high-speed cameras capturing 120 frames per second with a set of reflective markers on their face. We base our analysis on the locations of these marker points and create separate training sets of 24,000 frames for each of the two actors. We also select seven continuous conversations comprising of almost $152,000$ frames to form a testing set.

The IEMOCAP dataset provides discrete categorical labels as well as psychological data in terms of emotion attributes for each utterance, where an utterance is a sentence or similar period during which one actor talks continuously. Each utterance was labelled by the three expert human evaluators, who were allowed to assign more than one label to a single utterance in order to describe mixtures/blends of emotions, which are more common in natural communication. One of the limitations of this dataset is that the same label (or mixture of labels) was assigned to each frame in that utterance, based on the assumption that the emotional content did not change much within an utterance (duration ≈ 4.5 seconds). For the complete description of the dataset, see Section 4.6.

In this thesis, we will assume that the human evaluators are correct only where atleast two of them agreed and test the accuracy of our methods by comparing with the ground truth data of discrete labels as well as the values assigned to the continuous emotion attributes.

1.6 Main Research Contributions of Thesis

This thesis explores the field of affective computing focusing on emotion recognition, representation and analysis using a set of facial points as a clue. The research makes

contributions in terms of developing models using statistical and machine learning techniques, which are centred around three main research questions:

1. How can we map a set of facial points to basic as well as complex emotions?
2. Is it useful to represent facial changes over time in some emotion space?
3. What paths do emotions follow while moving from one state to another in an emotion space?

The following list described our contributions to the computational study of emotions:

1. We introduced a direct classification approach to map facial points to the basic emotion categories using shape models (Chapter 5), which can do the following:
   - describe a set of facial points as weighted basic emotions.
   - classify a set of facial points to a basic emotion category based on the minimum distance criterion.
2. We introduced an approach to represent facial points to an emotion space using a finite mixture model of von Mises probability distributions (Chapter 6), which can:
   - represent basic emotions at appropriate locations in that space.
   - represent complex emotions as mixture of basic emotions in that space.
   - estimate the direction of emotional expression, i.e., whether it is moving from positive to negative state or vice-versa.
   - estimate the change in intensity of emotional expression, i.e., whether it is moving from mild to intense state or vice-versa.

- develop a bridge between categorical and attribute-based descriptions of emotions by mapping emotions in terms of valence and activation using a set of matches to each basic emotion category.

3. We introduced ways to analyse paths of emotions on the emotion space by considering facial points representations as a time series (Chapter 7), which can do the following:

   - continuously track the paths of emotions on the emotion space through time.
   - test whether the emotion paths follow 'smooth' trajectories on the emotion space.
   - analyse the paths followed by emotions while moving from one state to another.

1.7 Overview of Thesis

This thesis is organised into 8 chapters (including this one). Chapter 2 will discuss psychological perspectives of emotions and Chapter 3 will provide a review of previous works related to automatic emotion recognition and analysis. Chapter 4 will give an overview of widely used emotion datasets and description of our selected dataset. Chapters 5, 6, and 7 will describe the main contributions of this thesis. The thesis will end with a discussion of the research in this thesis and will point out future directions in Chapter 8. An overview of each chapter is given below:

Chapter 2 will summarize some of the basic concepts and emotion theories from psychology which are essential for the computational modelling and analysis of emotions. The main purpose of this chapter is to build a bridge between psychological

theories and computational approaches to develop automatic systems for emotion recognition and analysis.

Chapter 3 will begin by describing two major computational approaches to automatic emotion recognition: the direct classification approach, and mapping to emotion space. The review of the literature will be conducted in order to develop an understanding of the different computational methods proposed by researchers from the machine learning, image processing, and computer vision community. This chapter will also point out gaps in the past research which implicitly explain why emotion recognition and analysis is in need of further study.

Chapter 4 will present criteria for selecting a suitable dataset for the task of emotion recognition and analysis, followed by a review of some of the widely-used existing datasets with respect to that criteria. The chapter will also provide an overview of the common data annotation tools used for discrete as well as continuous emotion labelling. On the basis of this survey, this chapter will describe our selected dataset and the associated benefits and limitations. The chapter will end with a description of how the data was preprocessed in order to identify informative facial points useful for the task of emotion recognition.

Chapter 5 will introduce a method for mapping facial points to the basic emotion categories using a set of shape models. This chapter will show that building separate shape models of different parts of the face and combining them can give more successful results than using only a model of the whole face. Statistical shape modelling based on Principal Component Analysis (PCA) will be described first, and then the classification technique will be presented. Detailed analysis of the shape models will be given, together with the experiments showing effectiveness of the proposed

method for recognising discrete basic emotions. The results of our algorithm will be tested by comparing it with state-of-the-art methods and we will demonstrate significant improvement in the reliability and performance accuracy. The chapter will end with a discussion pointing out the need for the recognition of complex emotions by identifying emotion blends in the data.

Chapter 6 will begin with a discussion of circular data analysis and will demonstrate why linear algebraic operations and standard statistical techniques cannot be used to analyse circular data. The chapter will address the problem of complex emotion recognition by using a degree of match to each basic emotion given by a set of shape models. A statistical modelling technique based on a mixture of von Mises probability distributions will be introduced, that combines each basic emotion in an appropriate amount for the recognition of complex emotions. The chapter will also provide a method for representing facial points movements on the activation-evaluation space, which is the most widely-used representation of emotions in psychological studies (described in Chapter 2). The success of the proposed technique will be demonstrated with experiments comparing the estimated results to those obtained by the psychological studies as well as to the ground truth data.

Chapter 7 will introduce computational methods to analyse the movement of facial points between different emotional states by considering the emotion paths representation in the activation-evaluation space as a time series. The chapter will also present the processes used to test the hypotheses listed in Section 1.3.2, together with the results of those tests. The chapter will end with a cross-validation test to evaluate the consistency of each of the proposed models and the reliability of the obtained results.

Chapter 8 will draw the conclusions and summarise the contributions of this thesis. The directions for future work will also be discussed at the end of this chapter.

Chapter 2

Emotions from the Perspective of Psychology

2.1 Introduction

While psychologists have been trying to define emotions for over 100 years, a non-subjective definition still eludes us [119, 171, 176, 184]. According to some theorists, emotion is an attribute of physiological functioning [49] and should be studied as a part of neuroscience, while others consider it to be a high level mental property [130, 131], and so a part of psychological studies. Another group merges the two concepts, which gives rise to a new approach called affective-neuroscience [167]. More recently, there have been significant efforts to recognise, represent, and analyse emotions based on their appearance, such as in the face, or tone of voice [144].

This chapter will summarize some of the basic concepts and emotion theories from psychology that are essential for the computational modelling and analysis of emotions. The main purpose of this chapter is to clarify the psychological terminologies

to be used in this thesis, and to provide background material for the development of automatic systems for emotion recognition and analysis.

2.2 Understanding Emotions

According to Barrett et al. [8], emotion is not a fundamental unit of mind, but rather a product of various simple components. These basic psychological components combine in different ways to produce a number of mental experiences, one of which is *emotion* [163]. These underlying components are referred to as 'ingredients' of emotion, and the essence of this process is termed psychological constructivism [124]. However, what these ingredients are and how they interact to produce different emotions is still not clear [8, 45]. In order to understand emotions and their psychological construction, we look at recent research in neuroscience based on the Iterative Reprocessing model of evaluation [47, 232]. This model suggests that it is not enough to know about the current feeling of a person; we must look at the dynamic temporal shifts in affect that lead him to his current state. For instance, a slightly positive emotional state will be considered as pleasant or aversive on the basis of whether the person was feeling worse or better in the recent past. This gives rise to a different emotion description for the same affective state [162, 187].

The Iterative Reprocessing model forms the basis of the Affective Trajectories Hypothesis in psychology, which suggests that emotions are dynamic states that are continuously updated based on the newest information, past information, and the prediction of what may happen next. These dynamic mental processes give rise to our 'affective trajectories' in time [45, 46, 123]. These trajectories provide a detailed description of how we reached our current state and what may be predicted next. This analysis suggests that by tracking emotion trajectories through time, we may

not only interpret the current emotion state more accurately, but may also predict the future states on the basis of emotion flow.

2.3 Basic and Complex/Mixed Emotions

There is a long discussion going back to Darwin [50] and Descartes [56] who support the existence of a small, fixed number of distinct emotions called 'universal basic emotions', such that all the people in the world express these emotions in the same way [130, 131]. In 1962, Silvan Tomkins [209] suggested that there are nine basic emotional states (one is neutral, two are positive and six are negative), each represented by a precise pattern of facial features. The discussion continues among many researchers [65, 85, 108, 160], who put forward their own list of basic emotions that differ in the type and the number of basic emotions proposed. In 1971, Ekman and Friesen [73,76,77] postulated six primary emotions (happiness, sadness, surprise, fear, anger, and disgust), each of which is claimed to be genetically determined and possesses a unique content together with a distinctive facial expression. These emotions are widely accepted as the basic emotions [86,179]. In 1990, Ortony and Turner [163] claimed that according to the psychological notion of basic emotions, frustration is also a basic emotion. In 1999, Ekman [68] further listed some possible basic emotions, referred to as the candidate basic emotions, which includes excitement, guilt, shame, and pride in achievement. Ortony and Turner [163] presented a list of basic emotions suggested by different researchers, some of which are given in Table 2.1.

In this thesis, we are using neutral state and a set of five emotional states (including basic emotions: happiness, sadness, anger as well as two candidate emotions: frustration and excitement) for our analysis. The reason of choosing this emotion set is two-fold, first, it includes similar emotions (e.g., happiness and excitement as well

Emotion Theorist	**Set of Basic Emotions**
James [111]	Fear, grief, love, rage
McDougall [145]	Anger, disgust, elation, fear, subjection, tender-emotion, wonder
Watson [219]	Fear, love, rage
Arnold [3]	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Mowrer [156]	Pain, pleasure
Izard [108]	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Plutchik [175]	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Ekman, Friesen, and Ellsworth [66]	Anger, disgust, fear, joy, sadness, surprise
Gray [89]	Rage and terror, anxiety, joy
Panksepp [166]	Expectancy, fear, rage, panic
Tomkins [210]	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Weiner and Graham [220]	Happiness, sadness
Frijda [85]	Desire, happiness, interest, surprise, wonder, sorrow
Oatley and Johnson-Laird [160]	Anger, disgust, anxiety, happiness, sadness

Table 2.1: List of basic emotions by various theorists. Presented by Ortony and Turner [163].

as anger and frustration) as well as opposite emotions (e.g., happiness and sadness) making it suitable to study relationship among various emotions. Second, for the missing emotions (disgust, surprise, and fear) there was insufficient data for complete analysis.

The research revealed that most human emotions are not pure examples of one basic emotion, but a mixture of them, known as complex emotions [50, 67, 109, 114, 146, 171, 176]. Basic emotions may be combined in various ways to give rise to several *blends* of emotions, e.g., contempt is a blend of disgust and anger and delight is a blend of happiness and surprise. In 1975, Ekman [74] pointed out thirty-three

different blends of six basic emotions which appear on the human face. He examined each of the six basic emotions separately and described how these emotions and their blends appear on the face. The blends may include more than two basic emotions, for example, anxiety is a mixture of fear, sadness, and anger [192]. In an emotion blend, the timing of *activation* of one emotion and *decay* of another may be different. For example, surprise is the shortest emotion, and decays quite quickly or blends with another emotion (mostly with sadness or happiness). It is possible that the two emotions are activated together, but one persists for longer [171].

2.4 Emotions and Facial Expressions

Facial expressions are highly modifiable and controllable, which brings into question the reliability of using facial expressions as a representation of underlying emotional states. The literature in psychology views two groups of emotion theorists in argument: one including Ekman [73], Izard [108], Malatesta [139], and Tomkins [209], who believe that the facial expressions represent the corresponding underlying emotional states. Others, like Fridlund [83], believe that we do not *express* but rather *display*, which does not necessarily have any relation with the felt emotions. In [70], Ekman presents the concept of 'display rules', which suggests that we gradually learn how to behave in public, what emotion to express and what to hide, when and how. There are some attempts at developing computational models to detect the posed, deceptive, or fake expressions using facial clues [9,33,134,188], but this is beyond the scope of this thesis.

In common with many researchers, we will assume that when we feel an emotion, it appears on our face [73,108,139,209]. Based on this, we will use facial expressions

as a measure of underlying emotions. This enables machine vision tools to be used to identify emotions, even though the emotions themselves are internal, private feelings.

2.5 The Space of Emotions

Over the past two decades there have been several efforts to define the structure of emotions in the field of psychology. One of the major assumptions is that the relations among emotions can be described in terms of their polarity and similarity to each other by a circular structure called a *circumplex*.

In 1941, Schlosberg [197] carried out an experiment in which he asked research participants to judge the emotions in 72 pictures showing different facial expressions. The participants were asked to categorise emotions into six categories: happiness, fear, anger, disgust, and contempt. Schlosberg found that the overlap of the judgments lead to a circular structure with two major dimensions depicting polarity, such as pleasant-unpleasant versus attention-rejection. In 1957, Block [22] asked a group of female students to assess 15 emotions using a 7-point bipolar scale having terms like good-bad, high-low, and active-passive. The correlation between ratings was factor analysed, which lead to two factors accounting for most of the variance. A circular order of emotions was obtained following the order: pride, anticipation, elation, love, contentment, sympathy, nostalgia, boredom, grief, guilt, humiliation, worry, envy, fear, and anger.

In 1980, Plutchik [174] used a paired-comparison method to rate the similarity of over 140 emotion words selected from the English dictionary. He used three different emotion words (accepting, angry, and sad) as *reference* words. Six judges were asked to rate over 140 words according to these reference words on a 11-point bipolar scale ranging from 'opposite' (-5), 'no relation' (0), to 'the same' (+5). The mean ratings

were converted to angular locations, such that if the mean rating of an emotion relative to the reference word had a plus sign, its angular location would be within 90° of the reference word, if the mean rating was 0, the angular position would be 90° away from the reference word, and if the mean rating had a minus sign, the angular location would be more than 90° away from the reference word. The resulting circumplex structure of emotions is shown in Fig. 2.1. It was found that similar emotions are found near one another (e.g., hostile and furious), while opposite emotions are almost 180° apart (e.g., pleased and unhappy). Some of the emotion words along with their angular placement are listed in Table B.1 in Appendix B. These measures of angular locations are also referred to as emotional orientation [44].

Several other methods were used to study the circular structure of emotions, including [173,174,190]. The differences in the sequence of emotions and their locations on the circumplex are partly related to the method used, and partly on the sample of emotions chosen for analysis.

2.5.1 Circumplex models of emotions

A circumplex model defines emotions in terms of their polarity and similarity to each other instead of placing them into discrete categories, thus providing a continuous description of emotions. In the literature describing the concept of the circumplex and its relation to emotions and personality, the theories presented by Plutchik and that of Russell seem to be the most prominent.

In 1980, Plutchik [174] presented a *psychoevolutionary theory* of emotions which suggests that all complex emotions are derivatives or mixtures of the basic emotions, in the same way as all colours are derived from the mixtures of three primary colours. On the basis of this similar property of colours and emotions, Plutchik suggests that an emotion circle may be created in the same way as a colour circle. An emotion

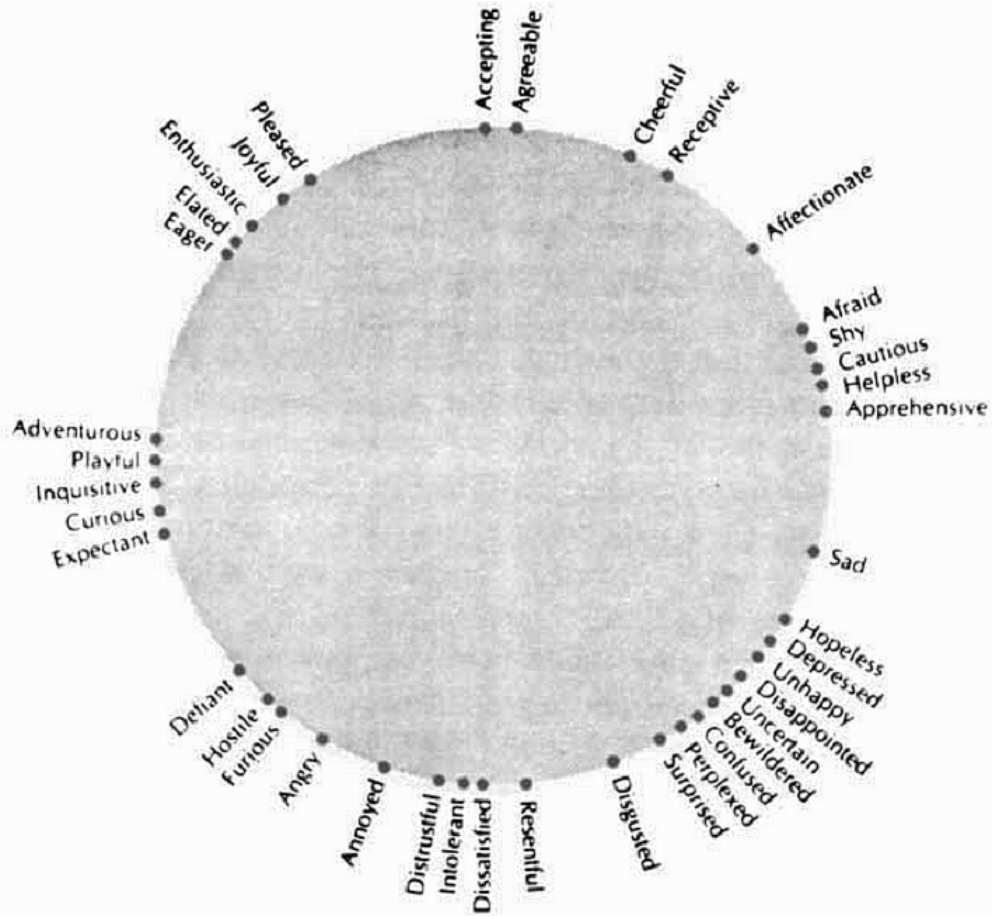


Figure 2.1: A circumplex structure of emotions obtained on the basis of similarity measures. Picture taken from [174].

circle or a circumplex is simply a reflection of certain relations among its elements, including *similarity* and *polarity* or *conflict*. Since emotions vary in degree of similarity to one another and show polarities (e.g., happiness versus sadness), it is possible to represent these relations by using a circular model. In terms of statistics, the correlation between emotions increases or decreases on the basis of their degree of similarity and of polarity. The *conflicting* emotions with -1 correlation correspond to the opposite poles of a circumplex (180° apart), independent or uncorrelated emotions with 0 correlation corresponds to a 90° divergence, while *similar* or positively correlated emotions lie close to each other on a circle.

In 1997, Russell [191] put forward certain properties to describe emotions, some of which are: (1) an emotion is not a standalone element, but rather a *member* of many categories, (2) the membership to each category is a matter of *degree* rather than all or none, (3) these categories are related to each other in a circular order, and (4) emotions vary in degree of *intensity* as a function of valence (positive or negative) and activation (active or passive).

By combining the theory and properties presented by Plutchik and Russell, we may conclude that complex emotions can be conceived of as mixtures of basic emotions, and the structure of emotions can be described by using a bipolar circumplex model.

2.5.2 Activation-evaluation space

The valence-activation (or activation-evaluation) space is the most widely-used bipolar circumplex model for representing emotions in psychological studies, and it is getting popular in computationally oriented research as well, e.g., [54, 106, 107]. A review of these and other related methods will be given in Chapter 3. This is a model that represents emotions based on their activation and valence/evaluation,

Activation refers to how motivated a person is during an emotional state. It is related to the degree of readiness to act, and differentiates between emotions having the same valence. For example, anger and sadness are both considered as negative emotions, but anger has high activation level than sadness, which differentiates between these two distinct emotions.

Evaluation refers to how positive or negative the emotion is. For instance, happiness is considered to be positive, while anger is considered to be a negative emotion based on their evaluation levels.

The validity of these dimensions has been agreed by several researchers, e.g., [53, 81, 164, 189, 195, 223]. In 1999, Lang et al. [128] developed International Affective

Picture System (IAPS) to provide a standarised set of emotionally-evocative photographs that can be used as a powerful emotional stimuli for experimental research related to emotions. The emotional judgments were based on two primary dimensions: valence, activation and a third, less strongly-related dimension: dominance which refers to how dominant or submissive an emotion. It was found that most of the variance in emotional assessments can be accounted for by two primary dimensions of valence and activation. There has been some efforts in defining the third dimension (including attention, potency, competence, or dominance) to explain the relation among emotions, but this has met with considerable disagreement [189, 222].

Activation-evaluation space forms a circular representation of emotions with neutral at the origin. The relative positions of emotions may be described by the specific angular locations on the circle based on their similarity to one another. For example, the angular location of anger is closer to frustration than to happiness. The radial distance from the centre of the circle represents the intensity of emotion. The greater the distance, the stronger the emotion and vice-versa [2, 177, 178]. Sometimes, the intensity of emotion is too slight to notice, or too subtle to effect what we do. Ekman [70] therefore suggested that we may better say that there are times when there is no emotion, i.e., the 'neutral' state. Neutral lies at the centre of the circle and is assumed to be the presence of any emotion with very low intensity. Fig. 2.2 shows a schematic representation of activation-evaluation space illustrating how emotions lie in a continuous order around a bipolar structure. The distribution of valence and activation states in this space is listed in Table 2.2.

One of the benefits of using the activation-evaluation space is that Whissell in [221] provided the corresponding values of valence and activation for a wide range of emotion words (including those selected by Plutchik in the paired-comparison study).

[1] activation (+/-), evaluation (+/-) near 0.

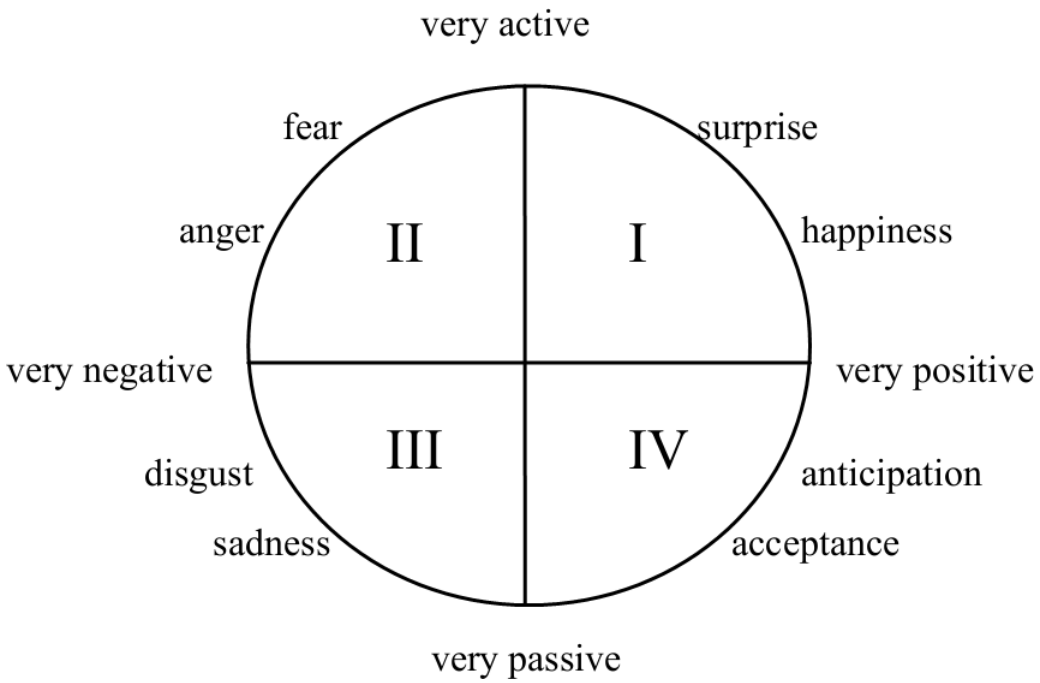


Figure 2.2: A bipolar circumplex of emotions, commonly known as activation-evaluation space. Picture taken from [234].

Label	**Activation/Evaluation States**
QI	positive activation, positive evaluation (+/+)
QII	positive activation, negative evaluation (+/-)
QIII	negative activation, negative evaluation (-/-)
QIV	negative activation, positive evaluation (-/+)
Neutral	close to the centre[1]

Table 2.2: Emotion classes in the activation-evaluation space [221]

They asked 33 students of psychology to assess almost 4,000 emotion words selected from the English dictionary and rate these words according to a 7-point scale for each dimension. For valence, 1 corresponds to 'very positive' and 7 means 'very negative'. Similarly, for activation, 1 represents 'very active' and 7 corresponds to the 'very passive' state. The mean ratings and standard deviations of ratings were calculated for each word along each dimension. There was significant correlation among the ratings of valence and activation for the entire group of words. Table B.1 in Appendix B

lists some of the emotion words and the corresponding values of valence and activation in the activation-evaluation space. A disadvantage of representing emotions in the activation-evaluation space is that some of the distinct emotions may be grouped quite close in that coordinate system and thus become indistinguishable, e.g., happiness and excitement both go along with high activation and positive valence and are difficult to identify as different emotions [194].

2.6 Categorical and Attribute-based Descriptions

Research reveals that human perception of facial expressions is categorical rather than based on valence and activation [52]. The words that we use to label and categorise expressions provide an insight into the overall structure of the underlying emotion [124]. Attribute-based labelling of expressions may lose a lot of such information. If we label an expression as *negative*, we do not get an idea about whether the underlying emotion is anger, fear, or disgust. Moreover, the so-called negative emotions may have positive functions, e.g., fear helps us avoid dangerous objects [69]. In contrast, amusement, which is assumed to be a positive emotion, may be negative, e.g., in the case of ridicule [70].

On the other hand, it is not possible to assign a categorical label to each expression of emotion. In this case, we normally assign two or more categorical labels to define an expression of a complex emotion. In this thesis, we will use both categorical and attribute-based description of emotions as complements to each other by describing complex emotions in terms of known basic emotions using activation-evaluation space. The attribute-based descriptions depict characteristics of each time-point in the space, while categorical labels may be assigned to a combination of underlying patterns followed by emotion paths in the activation-evaluation space.

2.7 Summary

This chapter has described the basic information about emotions from the psychological point of view. This can be used to start giving computers the abilities necessary to recognise emotions. It started with the discussion of the affective trajectories hypothesis, which helps to understand the dynamic nature of emotions. The concepts of basic and complex emotions were described. According to the psychoevolutionary theory, complex emotions may be conceived of as mixtures of known basic emotions, which gives a useful starting point to deal with a large number of complex emotional states.

Facial changes are the most obvious expressions of the underlying emotions that can be used by computer vision and image processing techniques to train automatic systems for emotion recognition. Further, it was found that the structure of emotions and their relations to one another can be described using a bipolar circumplex model. The chapter ends with a brief description of the categorical and attribute-based description of emotions, which suggests that both labels complement each other.

Chapter 3

Literature Review

3.1 Introduction

In the literature of psychology and computer science, there exists active research related to emotion recognition, representation, and analysis. The last chapter discussed the key concepts and models from psychology which are useful for enabling computers to recognise emotions. This chapter focuses on the review of computer vision and machine learning techniques commonly used for automatic emotion recognition and analysis.

There are two approaches to automatic emotion recognition: first, the direct classification approaches (based on the categorical descriptions of emotions) and second, mapping the emotions to quantised or continuous emotion spaces (based on the attribute-based descriptions of emotions). The literature reviewed in this chapter encompasses these two approaches. It includes methods using clues from facial expressions along with a brief description of the systems using multiple clues (e.g., facial expressions, voice signals, and body gestures). Both classification approaches have their own benefits and limitations. In this chapter, we will discuss some of the most

widely-used methods along with the state-of-the-art techniques for emotion recognition using both the direct classification approaches and mapping into emotion space approaches. There has been little work done on the detection and analysis of emotion dynamics; what there has been will be discussed briefly at the end of this chapter.

3.2 Direct Classification Approaches

The classification of emotions using facial expressions in terms of categories has been the most common approach in the field of automatic emotion recognition. Since the 1970s, various techniques have been proposed to categorise facial expressions into a set of basic emotions or a set of facial action units (the smallest visually distinguishable movements on the face [76]), using either static facial images or a sequence of facial images.

3.2.1 Template-based techniques

In this technique, a template is defined for each emotion category and the unknown facial expression is compared to all the defined templates. The category of the unknown facial expression is decided on the basis of the best matched template.

To classify unknown facial expressions to six basic emotions, Huang and Huang [104] applied the statistical shape model also known as the point distribution model (PDM) [38] to extract the facial features. They calculated 10 Action Parameters (APs) to describe the position variations of certain points on the facial features. PDMs are the models used for the analysis of shape variation marked with landmark points [39]. They applied Principal Component Analysis (PCA) [141] to the APs of all training expressions and found that 90% of AP variations are covered by the first 2 eigenvectors. PCA is a commonly-used dimensionality reduction technique to analyse sets

of data points in high dimensional spaces. This approach gives satisfactory classification accuracy, but the estimation of APs on the basis of gradient-descent-based shape parameters is a computationally expensive process. They reported an average computation time of seven minutes to analyse an image sequence approximately one second long. Also, the descriptions of the emotional expressions, defined in terms of facial actions, are incomplete.

Hong et al. [103] classified facial expressions into one of the six basic emotions or the expressionless (neutral) face using personalised galleries. It was assumed that two persons who look alike express emotions in a similar way. First, a labelled graph is fitted to the unknown face, then the best-matching person among those having the personalised templates is found by applying elastic graph matching [226]. The unknown facial expression is classified by using the personalised templates of the best-matched person. The system achieved 89% accuracy for the familiar faces (whose galleries were available), and 78% for unfamiliar faces. However, it only deals with full upright-frontal facial images and fails for profile or partially occluded facial images. The error rate of the system increases in the case of not finding the best-matched gallery, or the unavailability of the particular expression in the matched gallery.

Bartlett et al. [10] classified video frames into six basic emotions and neutral state by using Gabor energy filters [155], along with the recognition engines: AdaBoost [141], Support Vector Machines (SVM) [151], Linear Discriminant Analysis (LDA) [141], and feature selection techniques. They also classified frames to 18 action units individually or in combination, obtaining an accuracy of 94.5% on the Cohn-Kanade (CK) dataset [116]. Like [103], this system only deals with the frontal-view face images and does not account for the temporal dynamics (onset, apex, offset) of the facial expressions, or AUs.

In the work of Martins et al. [143], they used the Active Appearance Model (AAM) [37] to extract a facial geometry and Laplacian Eigen-Maps (LE) [18] to derive low-dimensional manifolds of that facial geometry. Martins et al. derived two types of manifolds: one which deals with the identity recognition, and the other for the person-specific facial expression recognition (expressions of six basic emotions and the neutral state). A multi-dimensional representation of a face can be represented by a single point in a multi-dimensional face-space and the variability of facial expressions can be represented as low-dimensional manifolds in that space [28]. The low-dimensional representation of facial changes in the face-space is a suitable approach to cover all possible variations of an emotional expression. However, the technique presented in this paper can only deal with the recognition of person-specific facial expressions of the familiar persons already present in the database and would fail for a face with unknown identity. Also, the 2D shape model ignores the detailed facial deformation which might improve the recognition of the minor variations of facial expressions.

Shan et al. [201] presented a person-independent facial expression recognition method to classify facial images into six basic emotions and the neutral state, based on the Boosted Local Binary Pattern (B-LBP) classifier. For details of the original LBP classifier, refer to [161]. However, their method only deals with static images without considering the temporal dynamics of facial expressions.

Bansal et al. [5] used Latent Dirichlet Allocation [21] along with the Hidden Markov Model to classify facial image sequences to six basic emotions. Generally, the technique of Latent Dirichlet Allocation is used in natural text processing. Using this technique, each image sequence is assumed to be a document and each frame in the sequence is a word of that document. In this way, a set of image sequences is represented as the set of topics assigned to each frame. A Hidden Markov Model for

each emotion was used to learn the sequence information of topics in image sequences, which is then used to classify image sequences to six basic emotion categories. However, the probability inference in LDA is a computationally complex process. When the number of emotion categories in a set of image sequence is small, the probability inference takes polynomial time, while in the case of several emotion transitions it is an NP-hard problem [204].

In [41], Costantini et al. demonstrated the role of upper and lower parts of the face in recognising emotions based on experiments including 74 participants. They compared the emotion recognition rate using the whole face, eyes, and mouth separately, and found that the eyes alone generate similar recognition rate to using the whole face, and higher recognition rate than using the mouth only. Several years ago, Bassili et al. [12] found that the upper part of face is associated with high recognition rate of negative emotions (e.g., sadness and fear), while the lower part of face yields high recognition accuracy of positive emotions (e.g., happiness). These results motivate the development of computational models of the full, upper and lower parts of the face separately, however in the reviewed literature we could not find any systematic experiments modelling and analysing the upper, lower, and full face templates separately for the task of emotion recognition.

3.2.2 FACS-based techniques

A very common method for measuring facial expressions in behavioural science is the Facial Action Coding System (FACS) [75, 76]. FACS is a scoring system defined for expert human observers, not computers. It aims to provide objective measures of facial activity to assist behavioural science analysis of the face. Ekman and Friesen defined 46 Action Units (AUs) that describe the smallest visually perceptible facial movements. They determined the effect of contraction of each of the facial muscles on

the visible appearance of face by using knowledge of facial anatomy. The system can be used to describe any facial movement (observed in images or videos) in terms of anatomically based action units. This system has been used, for instance, to demonstrate differences between genuine and simulated pain [135], differences between the facial signals of suicidal and non-suicidally depressed patients [101], and differences between when people are telling the truth versus lying [71].

A lot of researchers have tried to automate FACS by recognising facial action units in images and/or videos, e.g., Donato et al. [58] used Gabor wavelets and Independent Component Analysis (ICA) to recognise eight individual AUs and four combinations of AUs. Cohn et al. [35] used facial feature point tracking and discriminant function analysis to recognise eight individual AUs and seven combinations of AUs. Sixteen AUs were recognised by Tien et al. [208] using lip tracking, template matching, and neural networks in nearly-frontal facial images. Braathen et al. [24] recognised three AUs using particle filtering, Gabor wavelets, SVM, and HMM in facial images with varying head poses. Mahoor et al. [138] measured the intensity of AU12 and AU6 in videos captured from infant-mother interaction by using the SVM.

Most of the automatic systems using FACS are trained and tested on posed expressions where actors are asked to voluntarily contract specific muscles corresponding to certain AUs. Also, reliable FACS-coding of facial images/frames is a long tiring process which needs FACS-certified coders to spend hours to code just a few seconds of video. Although FACS is a promising approach, in reality it is not always possible to locate the action units in each image/frame due to changes in lighting conditions, occlusions caused by the facial hairs and glasses, and poor image quality. These problems also exist for other feature extraction techniques using facial images or videos.

3.2.3 Rule-based techniques

The rule-based techniques classify the unknown facial expressions to emotion categories based on rules applied to the movement of facial action units or the facial definition parameters (FDP) which define the shape of the face [165].

Pantic et al. [168] defined rules based on FACS to classify facial expressions to 31 action units as well as six basic emotions. The rules were applied to the model deformation parameters calculated by taking the difference between the detected model features and the same features detected in the neutral face of the same person. The system reported high average accuracy, but is limited to just static images of posed expressions.

Zhou et al. [236] divided the facial expressions into three categories based on the deformation of mouth region. Anger, sadness and disgust come in the first category, happiness and fear in the second category, and surprise in the third category. Classification to the six basic emotions was done by applying rules to the displacement of the key points of eyes, eyebrows, and mouth extracted by using an Active Appearance Model. As in [168], this system has also been tested on static facial images of posed expressions without considering the temporal dynamics of facial expressions.

One of the limitations of rule-based systems is that the rules are only applied properly if the feature set is extracted correctly from images. The reliable extraction of facial features is a difficult task for images which are captured under complex backgrounds, varying lighting conditions, showing spontaneous expressions, uncontrolled head movements, and/or are partially or fully occluded.

The direct classification of emotions in terms of discrete categories has been the focus of researchers, but it fails to describe the wide range of emotions that occur in daily communication and also it ignores the varying intensity of emotions. There are some efforts to detect non-basic emotional states from deliberately posed facial

expressions, e.g., fatigue [113], and some mental states such as concentrating, interested, thinking, and confused [118, 229]. In any case, the categorical approach just presents a list of discrete emotions without exhibiting any link between them. Every emotion is considered as an independent state with no relationship to other emotional states. In the following section, we review some state-of-the-art methods for automatic emotion recognition by mapping emotions to emotion spaces.

3.3 Mapping to Emotion Space Approaches

After decades of research in emotions, researchers have shown that in everyday interactions people exhibit non-basic, intermediate, or complex emotional states which are related to each other in a systematic manner. Based on this, it is not appropriate to assign a single independent categorical label to complex emotions which are mixtures of more than one emotion category. Therefore, there is a shift in research towards emotion space representations, where emotions are mapped into an emotion space either in quantised levels or along a continuum, in part to recognise the fact that emotions are a continuous phenomenon and in part to enable complex emotions to be identified without requiring labels [92].

3.3.1 Quantised approaches

In this approach, the emotion attributes (e.g., valence and activation) are quantised into an arbitrary number of levels or intensities. In this approach the most common way is to reduce the emotion classification problem to a two-class problem (positive versus negative and active versus passive) or to a four-class problem (quadrants of 2D activation-evaluation space).

Ioannou et al. [107] classified emotions in video frames to the neutral state, the six basic emotions as well as three quadrants of the activation-evaluation space (there was no emotion lying in the fourth quadrant). They defined fuzzy rules based on the variations of Facial Action Parameters (FAP). FAPs were defined by Pandzic et al. [165], and are closely related to muscle actions, and represent a complete set of basic facial actions along with head motion as well as eye, tongue, and mouth control. The system is dependent on the robust extraction of FAP variations, which is quite difficult in the case of naturalistic data. Also, the classification to the quadrants of the activation-evaluation space does not give much information about the emotional state, since each of them contains emotions expressed with highly varying features (e.g., anger and fear both lie in the same negative/active quadrant) [120].

Shin et al. [202] used facial expressions as a clue from the Korean facial expression database [4] selecting 252 static images for training and 66 images for testing. The images consists of 11 expressions of 6 subjects (3 males and 3 females). They classified expressions into 9-point scale for valence and activation using manifold learning for the feature extraction of facial expressions. The system reported 90.9% accuracy for valence and 56.1% for activation. However, the evaluation has been done on a very small testing set, which raises doubts about the reliability of the reported performance accuracy.

Following the quantised approach of emotion classification, most of the published work in the literature uses multiple clues, e.g., Karpouzis et al. [120] used clues from facial and hand gestures as well as vocal information to classify naturalistic emotions into neutral state and 3 quadrants of the activation-evaluation space. Similarly, Caridakis et al. [26] discriminated emotions into 5 classes (neutral state and four quadrants of activation-evaluation space) using a feed-forward back-propagation network. The system performed decision-level fusion of two visual modalities, i.e., facial

expressions as well as hand and body gestures using the Sensitive Artificial Listener (SAL) dataset [60]. As in [107], these two approaches also classify emotions to the quadrants of the activation-evaluation space, which makes it difficult to differentiate between emotions within the same quadrant.

Kulic et al. [126] used facial muscle contraction, heart beat, and perspiration rate to classify emotions into low, medium, and high levels of valence and activation. They implemented a Hidden Markov Model to estimate emotion attributes of 36 human subjects during human-robot interaction. The system is based on the physiological signals gathered by using different sensors on the body, which makes it invasive and unsuitable for capturing naturalistic emotion data.

Some efforts have been made on emotion classification based on quantised approaches using motion capture signals. For example, Wöllmer et al. [228] used the facial markers information as well as audio clues from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [25] to classify a set of utterances to three levels (negative, neutral, and positive) in terms of two emotion attributes: valence and activation. They used Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Networks (RNN) which consists of two neural networks: the first processes the sequences forwards and the other processes it backwards. In this way, the BLSTM RNNs have access to past and future data points in the sequence. However, their system is based on the segmented utterances (not continuous conversations), which often contains long breaks due to several silent frames in a sequence. The performance may be improved by utilising this missing data while considering the contextual information.

The quantised approach to emotion recognition is a step forward to the direct classification approaches in a sense that it gives more information about the emotion attributes and intensity. Also, it enables the classification of relatively large numbers

of discrete emotions as compared to just the basic emotions or a few action units. On the other hand, it is a simplified problem in terms of emotion space approaches that can represent a set of discrete emotions in terms of emotion attributes. Although it gives an overall idea about the nature of emotional states by placing them into quadrants or assigning some level of intensity, it still leaves them indistinguishable in terms of fine categories. Nevertheless, neither of these two approaches is able to analyse the dynamics of a large number of natural emotions and their relationships to each other. Moving from one basic emotion to another in the case of categorical description and from one quadrant to another in the case of quantised labelling would not make much sense in real life scenarios. In the following section, we describe the continuous approaches for emotion recognition by mapping them into emotion spaces.

3.3.2 Continuous approaches

In real life, we come across continuous complex emotions rather than discrete basic categories or quantised labellings. The term 'continuous' refers to the uninterrupted sequence of emotions, which are dynamic in terms of the changing facial patterns, the speed of onset, apex and offset movements, their intensity and duration. There are fuzzy boundaries between emotional states which are too vague to be separated. Based on this, research in computer vision is shifting towards continuous emotion recognition by mapping emotions into the emotion space continuum to explore the dynamics of complex emotional states. However, the research in this area is still in its infancy.

In 2008, Kanluan et al. [117] used facial expression and audio clues to classify emotions in terms of three continuous emotion attributes (valence, activation, and dominance). They used SVM Regression (SVR), k-Nearest Neighbor estimator and

a fuzzy logic estimator to estimate emotion attributes using both modalities separately. The emotion estimation for each modality was fused at a decision-level using a weighted linear combination. The system used the VAM corpus [90] (videos recorded at 25 frames per second) in which the data annotation was done on a 5-point scale (mapped to a scale of [-1, +1]) for each emotion attribute. The results show that valence was best estimated using the visual clues, while the estimation of activation and dominance was best using the combination of visual and audio information. These results might be influenced by the choice of facial features used in the training, i.e., eyes and lips. The lower face (i.e., lips/mouth) region is not considered as a good estimator of emotions, especially in *talking* scenarios as in the chosen dataset [41].

Hupont et al. [106] in 2010 presented a system for mapping complex emotions to the activation-evaluation space based on the positions of discrete basic emotions and the neutral state in that space. The corresponding angles, valence, and activation values of basic emotions and the neutral state are listed by Whissell in [221]. The basic emotions were classified based on the distances and angles between a set of facial points. This was done by using a combination of classifiers implemented by the Waikato Environment for Knowledge Analysis (WEKA) tool [100]. The unknown emotion was mapped to the activation-evaluation space using a weighted linear combination of the position (in x, y coordinates) of the basic emotions in that space, although it is unclear from their paper how this was actually done. However, it is somehow based on the valence, activation, and angular values listed by Whissell. The weights are associated with the confidence value of classifying an unknown facial expression using a set of basic emotion classifiers. The authors reported that it is possible to measure the intensity of emotions based on their attributes (valence and activation) values, but do not say how.

In 2011, Nicolaou et al. [158] used the Output-Associative Relevance Vector Machine (OA-RVM) regression framework to predict valence and activation from naturalistic facial expressions. The presented framework is based on learning non-linear input and output dependencies in the emotion data. The system was tested on the SAL dataset reporting high prediction accuracy as compared to both RVM and SVM. The proposed framework is quite robust for continuous emotion classification to valence and activation, however further analysis of temporal dynamics is needed in order to understand the correlation between these two emotion attributes.

In the work of Dahmane et al. [48] in 2011, they used Gabor energy filters to extract facial features and multi-class SVM to classify emotion in terms of four attributes: valence, activation, expectancy, and power. The results show quite low recognition accuracy for valence (48%) and power (36%), reaching an overall accuracy of 51.6% for all emotion attributes using the SEMAINE dataset [147]. The classification was done by sampling video frames at an interval of 10 frames (videos were recorded at 50 frames per second) which might lose some information.

In 2012, Martinez et al. [142] presented a model (without actual implementation) consisting of C distinct continuous spaces one for each basic emotion category. The spaces are linearly combined to represent *blends* of emotions. The intensity of each contributing emotion in the blend defines the weight of each emotion in the combination. The idea of representing complex emotions by mixing the basic emotions is similar to our approach for complex emotions recognition, but the linear combination of emotion spaces proposed in this model is not suitable to characterise complex emotions. The reason why linear operations are inappropriate to analyse emotion spaces will be illustrated in Section 6.2.

Some efforts have emerged for continuous emotion recognition using audio modality, e.g., [90, 132, 227], head gestures, e.g., [93], and thermal signals, e.g., [149, 157].

However, to the extent of our knowledge the continuous emotion recognition and analysis has not been attempted yet using motion capture data.

3.4 Analysis of Emotion Dynamics

Temporal dynamics of emotions play an important role in the proper interpretation and understanding of emotions [196]. The information about facial expressions through time helps to interpret the relationships between emotions, such as how the intensity of one emotion changes while transitioning from one state to another, and what paths the emotions follow during emotion transitions. In the literature, very few efforts have gone into detecting and analysing temporal dynamics of emotions, focusing only on detecting whether a certain facial expression or a combination of AUs is in its onset, apex, or offset phase.

In the work of Valstar et al. [213], they detected the presence of any of 15 AUs per frame along with some aspects of their temporal dynamics, i.e., whether the detected action unit is in its onset, apex, or offset phase and the total duration of activation of that AU. They based this analysis on the tracking data of 20 facial points detected using the GentleBoost [84] template. The action units and their temporal dynamics were classified using SVM on the posed video sequences from the MMI dataset. This system is helpful to study the movements of some of the facial muscles, but is of less help for facial expression classification. One of the reasons is that the number of AUs detected by this system is too small to be mapped to a wide range of facial expressions. Also, the system is evaluated on the posed neutral-onset-apex-offset-neutral sequences of AUs, that makes it unsuitable for practical applications.

Gunes et al. [95] detected the emotion segments by finding the start and end of the neutral-onset-apex-offset-neutral phase from face and body videos by comparing each

frame to the reference (neutral) frame as well as consecutive frames. Using the apex frame out of the detected emotion segment, they detected six basic emotions as well as boredom, anxiety, uncertainty, puzzlement, and surprise (positive, neutral, and negative) using facial expressions and body gestures. The system was tested on the Bimodal Face and Body Gesture (FABO) [94] dataset, where subjects were asked to act out certain emotion-eliciting scenarios in laboratory settings. The system detects the apex from the posed neutral-onset-apex-offset-neutral sequence of emotions, which has limited its use in naturalistic scenarios.

To the best of our knowledge, so far none of the computational studies has focused on the analysis of temporal dynamics in order to understand the relationships between different emotions, the paths followed by emotions while moving from one state to another (i.e., emotion trajectories), or to study the change in emotion intensity through time.

3.5 Summary

This chapter has demonstrated two main approaches to automatic human emotion recognition: the direct classification approach, and mapping to emotion space. The direct classification approaches include classification to a set of discrete emotion categories or facial action units. Most of the published work using this approach has focused on recognising the posed basic emotions recorded in strict laboratory settings. There are several techniques for recognising emotion categories or facial action units using audio-visual clues including facial expressions, voice, and body gestures. In this chapter, we have discussed some of the widely-used methods broadly categorised into template-based techniques, FACS-based techniques, and the rule-based techniques for classifying emotions to discrete categories.

The literature shows that the classification techniques for basic emotion categories reported satisfactory results. However, the research in psychology and related fields have shown that our real life emotions not only include basic emotions, but a wide range of complex emotions (mixtures of more than one basic emotion). It seems inappropriate to assign a single categorical label to such complex emotional states, therefore the research is shifting to classify emotions by mapping them into emotion spaces. In these approaches, emotions are represented along a continuum (e.g., from -1 to $+1$) in terms of emotion attributes (e.g., valence and activation) instead of discrete categories.

This chapter has reviewed a range of techniques for mapping emotions to emotion spaces, which are broadly categorised into quantised and continuous approaches. There is relatively more work done in emotion recognition using quanitised approaches than the continuous mapping of emotions to emotion spaces. Among the emotion spaces used for representing complex emotions, the activation-evaluation space is the most widely used. Very few methods in the literature focused on the analysis of temporal dynamics of continuous emotions, and to the best of our knowledge there is no computational study about the analysis of emotion trajectories on the emotion spaces.

The literature has demonstrated that despite several efforts in the field of automatic emotion recognition, still there are a lot of challenges and this field deserves further attention. In the next chapter, we will present a review of existing datasets commonly used for human emotion analysis and describe the widely-used data annotation tools. The next chapter will also discuss our chosen dataset, along with its benefits and limitations.

Chapter 4

Datasets for Emotion Recognition and Analysis

4.1 Introduction

In the literature, most of the emotion datasets are collected using one of the three main techniques: in the *first*, subjects are asked to pose the given emotions in strict laboratory settings; in the *second*, subjects are asked to act out 'induced' emotions as naturally as possible in simulated natural settings; and in the *third*, natural interactions between subjects, e.g., television talk shows, interviews, or debates are recorded.

The posed emotions are quite different from natural real-life emotions in duration, intensity, and facial muscle actions. For instance, natural smiles are of smaller amplitude, low intensity, shorter in duration, and cause the contraction of orbicularis oculi (muscles around the eyes) which is not voluntarily possible in the case of posed smiles [72]. Using the second method, the naturalistic-acted emotion data is collected, which depends strongly on the actor's skills. Some skilled actors can act out emotions

very much similar to the natural emotions. The benefits of using this method is to get 'natural-like' (close to natural) data under controlled recording environments (including settings of cameras and/or microphone) and to avoid out-of-the-plane movements of head and body, occlusions, dynamic background, and varying lighting conditions. The third method provides ideal data for natural emotions, but is difficult to use in training due to the associated challenges of uncontrolled environment and body movements. Poor training, lack of control, and unreliable annotation of natural data leads to erroneous results [59].

In this chapter, we will present our criteria for selecting a dataset for emotion recognition and analysis followed by the review of some of the widely used existing datasets with respect to that criteria. The chapter also gives an overview of the common data annotation tools used for discrete as well as continuous emotions labelling. Based on the survey of existing datasets, we will describe our selected dataset and the associated benefits and limitations. The chapter ends with the description of how the data was preprocessed.

4.2 Criteria for Selecting a Dataset

We set some criteria for the selection of a dataset suitable for emotion recognition and analysis as follows:

1. The size of dataset should be large enough to judge the reliability of the system.

2. The data should be collected under a controlled environment to ensure reliable extraction of features for training.

3. The emotion data should be as natural as possible to enable real-life emotion recognition.

4. The dataset should contain motion capture data of facial points instead of full facial images to avoid preprocessing steps such as face detection and facial features extraction.

5. The dataset should describe emotions in terms of discrete categories as well as attribute-based labelling (e.g., valence, activation). These two approaches complement each other and improve the reliability of data.

6. The dataset should be annotated continuously (ideally frame-by-frame) through time.

4.3 Review of Datasets

Table 4.1 presents a review of some of the noteworthy visual and audiovisual data resources in chronological order as reported in the published literature. The table presents the elicitation methods (posed, induced, or natural), size (number of subjects participated and number of available images/videos), type (visual (V), or audiovisual (AV)), emotion description (categorical labels, FACS coding, or continuous annotation in terms of emotion attributes, e.g., valence and activation), and the criteria (as listed in Section 4.2) met by the datasets. For a more detailed review of the datasets, please refer to [63].

References	**Elicitation Methods**	**Size**	**V/AV**	**Emotion Description**	**Criteria Met**
Cohn-Kanade(CK) dataset, 2000 [116]	Posed	210 subjects, Available: 486 videos	V	FACS coded and 6 basic emotions	1, 2
The Belfast Naturalistic Database, 2000 [61]	Natural	100 subjects, 239 clips (209 from TV Recordings, 30 from interview recordings)	AV	Valence, Activation, Intensity. Categorical labelling by selecting words from two lists: one containing 16 coarse emotion words and other containing 24 fine-grained emotion words	1, 3, 5
Joint processing of audio-visual information for the recognition of emotional expressions, 2000 [29]	Posed	100 subjects, 9900 visual and AV expressions	AV	6 basic emotions, interest, puzzle, boredom, and frustration	1, 2

References	Elicitation Methods	Size	V/AV	Emotion Description	Criteria Met
The Emotional Integration of Childhood Experience during the Adult Attachment Interview, 2004 [186]	Natural	60 subjects. Each interview lasts 30-60 minutes	V	FACS coded, positive, and negative	1, 3
Database of Moving Faces and People, 2005 [159]	Natural	229 subjects	V	6 basic emotions, puzzle, laughter, boredom, disbelief, and others	1, 3
EmoTV1 database, 2005 [1]	Natural	51 video clips recorded from French TV channels	AV	Anger, despair, disgust, doubt, exaltation, fear, irritation, joy, neutral, pain, sadness, serenity, surprise and worry. Valence, and activation	1, 3, 5

References	**Elicitation Methods**	**Size**	**V/AV**	**Emotion Description**	**Criteria Met**
The SAFE Corpus, 2005 [32]	Natural	400 audio-visual sequences from 8sec - 5min extracted from 30 recent movies on DVD support	AV	Fear including sub-categories (stress, terror, anxiety, etc.). Positive, neutral, and negative	1, 3
A 3D Facial Expression Database For Facial Behaviour Research, 2006 [230]	Posed	100 subjects	V	6 basic emotions and 4 levels of emotion intensity (low, middle, high, and highest)	1, 2
A Bimodal Face and Body Gesture (FABO) database, 2006 [94]	Posed	23 subjects, Available: 210 videos	V	6 basic emotions, neutral, uncertainty, anxiety, boredom	1, 2
RU-FACS, 2006 [11]	Induced	100 subjects	V	FACS coded (33 AUs)	1, 2, 3, 6

References	Elicitation Methods	Size	V/AV	Emotion Description	Criteria Met
The HUMAINE database, 2007 [62]	Natural and Induced	50 clips (taken from 3 natural and 6 induced datasets)	AV	Intensity, valence, activation, power, expectation, simulating, masking, and one of the 48 emotion words	1, 2, 3, 5, 6
Authentic Facial Expression Database, 2007 [200]	Natural	28 adults	V	Neutral, happy, surprise, disgust	3, 6
Interactive Emotional Dyadic Motion Capture Database (IEMOCAP), 2008 [25]	Induced	10 subjects (5 males and 5 females), 12 hours data (dyadic conversations with markers on the face)	AV	Angry, happy, sad, neutral, frustrated, surprised, fearful, excited, disgusted, valence, activation, and dominance	1, 2, 3, 4, 5

References	**Elicitation Methods**	**Size**	**V/AV**	**Emotion Description**	**Criteria Met**
The Sensitive Artificial Listener (SAL) for generating emotionally coloured conversation, 2008 [60]	Induced	24 adults, 10 hours of audio-visual recordings	AV	Valence and activation. A part of SAL database (SAL 1) has been labelled in more detailed way as a part of the HUMAINE dataset [62]	1, 2, 3, 6
The Vera am Mittag german audio-visual emotional speech database, 2008 [90]	Natural	20 participants, 12 hours of audio-visual recordings of TV talk show	AV	6 basic emotions. Valence, activation, and dominance	1, 3, 5, 6

References	Elicitation Methods	Size	V/AV	Emotion Description	Criteria Met
The Montreal Affective Voices (MAV) Database, 2008 [17]	Induced	90 emotionally-coloured pronunciations of the word 'ah' by 5 male and 5 female actors	AV	Anger, disgust, sadness, fear, pain, happiness, pleasure, surprise, neutral. Valence, activation, and intensity	1, 2, 3, 5, 6
The SEMAINE corpus of emotionally coloured character interactions, 2010 [147]	Induced	20 participants: 100 conversational and 50 non-conversational recordings	AV	Valence, activation, power, expectation, intensity, and 6 basic emotions	1, 2, 3, 5, 6

References	**Elicitation Methods**	**Size**	**V/AV**	**Emotion Description**	**Criteria Met**
Extended Cohn-Kanade (CK+), 2010 [136]	Posed basic emotions and Natural smiles	123 subjects, Available: 590 videos	V	FACS coded and 6 basic emotions	1, 2
MMI Facial Expression Extended Database, 2010 [169, 214]	Posed and Induced	1250 videos, 600 static images	V	FACS coded (31 AUs), 6 basic emotions, and some complex emotions	1, 2, 3, 6
The Natural Visible and Infrared facial Expression (NVIE) database, 2010 [217]	Induced and Natural	almost 100 subjects captured by natural and infrared cameras	V	Disgust, fear, happiness. 4 quadrants of activation-evaluation space $(++, +-, -+, --)$	1, 2, 3
The Geneva Multimodal Emotion Portrayal (GEMEP) Corpus, 2010 [6]	Induced	7000 emotions portrayals by 10 professional actors	AV	FACS coded	1, 2, 3

References	**Elicitation Methods**	**Size**	**V/AV**	**Emotion Description**	**Criteria Met**
The Belfast Induced Natural Emotion Database, 2012 [203]	Induced	137 male and 189 female subjects	AV	Valence, activation, and intensity of basic emotions	1, 2, 3, 5
A High-Resolution Spontaneous 3D Dynamic Facial Expression Database, 2013 [235]	Induced	41 participants, 328 sequences (from onset to offset)	V	FACS coded (27 AUs) by 2 coders, self-reported emotions using 5-point Likert scale (relaxed, happiness, disgust, fear, anger, sadness, sympathy, surprise, startle, physical pain, and embarrassment)	1, 2, 3, 6 (frame-level ground truth for facial actions only).

Table 4.1: A review of existing visual (V) and audio-visual (AV) datasets for human emotion recognition and analysis with respect to the presented criteria.

4.4 Data Annotation Tools

Generally, data annotation for each modality (e.g., voice, facial expressions, body gestures, and thermal signals) is done separately assuming the independence between the channels. The most common tool for annotating audio (or audio-visual) data is FeelTrace [43]. FeelTrace is based on the activation-evaluation space and allows the observers to track emotions in speech as they perceive it. The tool represents space as a circle on the computer screen and the observers move the mouse cursor to the appropriate point on the circle to label emotional states. The movement of cursor is colour-coded and key emotions are located as 'landmarks' at certain points on the space (as shown in Fig. 4.1). The location of landmarks was based on the BEEVer (Basic English Emotion Vocabulary) study [42]. FeelTrace provides fast and reliable annotation of continuous emotions but lacks cross-platform support and has restrictions on the integrated video player functionality [102].

Devillers et al. [57] presented the Multi-level Emotion and Context Annotation Scheme (MECAS) addressing the challenges associated with the annotation of real-life non-basic/mixed emotions. The labelling was done using a modified Transcriber tool [7] which enables the labellers to select two labels: 'major' and 'minor' for the mixed emotions. The labels are organized into 3 level emotion-class hierarchy from coarse-grained (7 emotion classes: fear, anger, sadness, hurt, surprise, positive, and neutral) to fine-grained (21 emotion classes including neutral) labels. Among the coarse-grained emotions; fear, anger, sadness, hurt are grouped to the 'negative' valence-level and surprise comes under 'negative or positive' valence-level. This tool is limited to annotate audio data only providing segment-level emotion annotation in terms of emotion classes. It does not provide continuous labels in terms of emotion attributes (e.g., valence and activation).

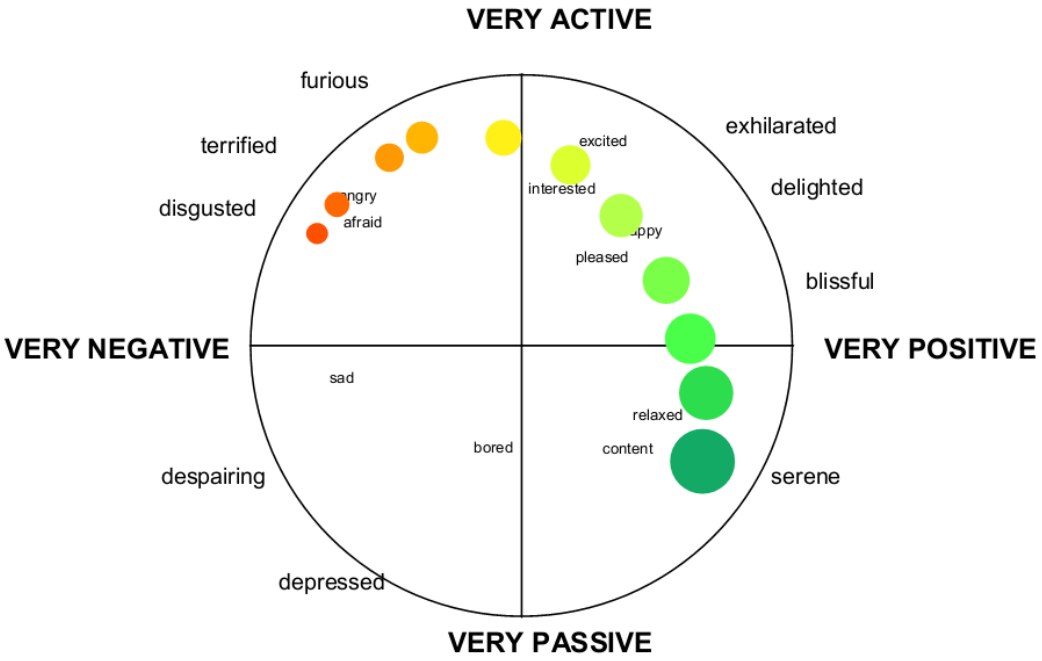


Figure 4.1: Example of FeelTrace display during a tracking session. Cursor colour changes from red/orange at the left hand end of the arc, to yellow beside the active/passive axis, to bright green on the negative/positive axis, to blue-green at the right hand end of the arc [43].

A widely-used freely-available research tool for annotating audiovisual data is "*Annotation of Video and Spoken Language*" (ANVIL) [122], which is a generic multi-layered approach containing time-anchored elements. Each 'element' contains multiple attribute-value pairs. The 'layers' (also called tracks) include words, dialogue, gestures, and postures. If an element has a start and end time it is called a 'primary' element, e.g., a word, however an element containing multiple primary elements is referred to as a 'secondary' element, e.g., a dialogue. ANVIL supports 'attribute' types including string, boolean, list of strings, and pointers, each defining a range of possible 'values'. ANVIL is platform-independent, fast, and easy to use tool and has been used extensively for emotion annotation purposes, e.g., [25, 31, 62]. However, it needs improvement in terms of output formats and the interface [122].

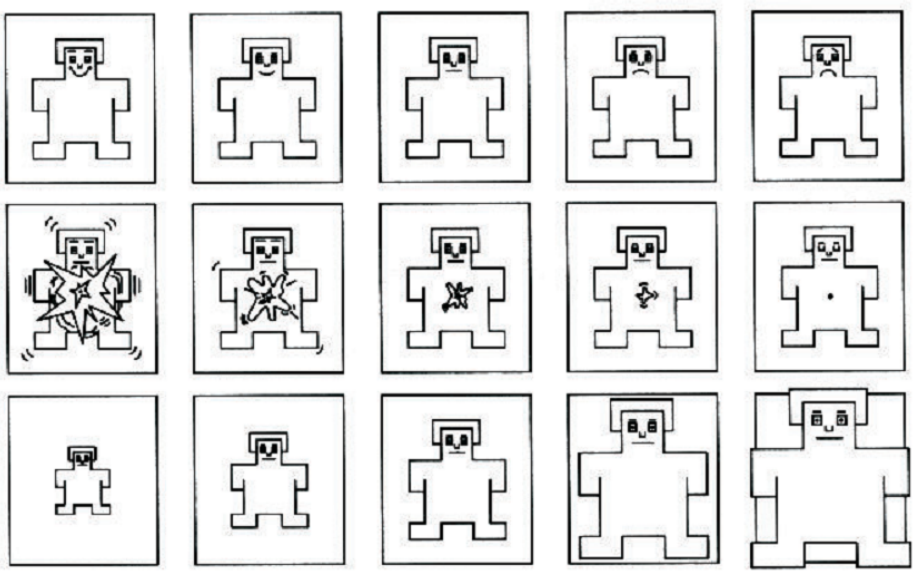


Figure 4.2: Iconic representation of valence (top row), activation (middle row), and dominance (bottom row) by Self Assessment Manikins [127].

A simple yet efficient tool called Self Assessment Manikins (SAMs) was proposed to label emotions in natural speech [127]. SAMs used the iconic representation of the three affective attributes: valence, activation, and dominance which refers to how dominant or submissive an emotion is, e.g., both anger and fear are negative emotions but anger is a dominant emotion, while fear is a submissive emotion. Fig. 4.2 shows a representation of SAMs. For each SAM, the selection can be mapped to an integer from 1 to 5. The presented technique was tested on utterance-based emotion segments, reporting a high inter-evaluator agreement. The limitation of using SAMs is associated with its representation of affective attributes. It is difficult to finely differentiate between the emotions having almost the same valence and activation values, e.g., anger and fear.

'Emotion slider' is another tool for annotating the emotions based on self-reported valence information from subjects interacting with a system [129]. As the annotation is solely based on valence, it is difficult to differentiate between emotions having the same valence values.

4.5 The Selected Dataset

In the list of existing datasets presented in Table 4.1, there are five datasets that met most of the criteria listed in Section 4.2. These are the HUMAINE dataset, the Montreal Affective Voices Dataset, the MMI Facial Expression Extended dataset, the SEMAINE dataset, and the IEMOCAP dataset.

The HUMAINE dataset contains data from six induced emotions datasets: the Sensitive Artificial Listener (SAL) dataset, Spaghetti dataset, Belfast Driving simulator dataset, EmoTABOO dataset, Green Persuasive dataset, and DRIVAWORK dataset. Out of these, the SAL dataset records interactions between a human user and an agent (which is, or appears to be, a machine). The agent acts out four different personalities (i.e., happy, gloomy, angry, and pragmatic), and the user chooses from which personality he wants to talk to. The dataset produced a large range of emotions, but except 'happiness' the emotions are too subtle to be detected accurately. On the other hand, the Spaghetti dataset contains quite intense basic emotions induced by an activity where the subjects were asked to feel inside a box containing a warm bowl of spaghetti. The subjects did not know in advance what is in the box, which leads to quite intense responses. However, the dataset is too short to enable reliable training of emotion recognition system and the real-life emotions are generally not as intense as in this dataset. The rest of the datasets are recorded in driving and gaming environments, which do not fulfill the requirement of natural communication.

The Montreal Affective Voices dataset contains AV recording of emotionally coloured pronunciation of the word 'ah'. Like the Spaghetti dataset, the emotional segments are too short to cover continuous contextual information, making it difficult to analyse emotion dynamics through time.

The MMI Facial Expression Extended dataset contains posed as well as induced emotions data of six basic emotions and some complex emotions. The dataset provides

frame-level FACS coding indicating whether the frame is in its neutral, onset, apex, or offset phase. While this dataset is very suitable for analysing facial expressions in terms of action units, due to the lack of attribute-based labelling of emotions it is less suitable for training a system to recognise complex emotions by mapping them into emotion spaces. Similarly, the spontaneous 3D dynamic facial expression dataset contains manual coding of 27 action units, but does not contain emotions attribute-based labelling.

The SEMAINE and IEMOCAP datasets are suitable for the task of emotion recognition and analysis. However, the SEMAINE dataset is developed quite recently and was not available at the time of dataset selection for our research. We chose to use IEMOCAP to demonstrate the efficacy of our algorithms. The selected dataset fulfills most of the criteria for the dataset selection (listed in Section 4.2) with the exception of continuous annotation of facial expressions. Assuming that the emotion would remain same within one utterance (average duration 4.5 seconds), the same label was assigned to each frame in that utterance.

IEMOCAP contains motion capture data of the markers attached to the face of subjects during recording. The markers' layout is based on the Feature Points (FPs) defined in the MPEG-4 Facial Animation (FA) standards [165]. The MPEG-4 FA standard defines 84 FPs on the morphological places of the neutral head model. These feature points cover all of the basic facial characteristics and are considered to be the best way of representing a human face. For this reason, the selected dataset is an acceptable alternative to using the full facial images for analysis.

The use of markers seems to be a constraint in recording natural human interactions, but actually the size of markers was quite small (diameter $\approx$ 1 cm) and the actors reported no problem in communicating naturally during the session [25]. Using markers is a good starting point for the systems focusing on emotion analysis in order

to minimise preprocessing steps associated with the face detection and facial feature extraction techniques.

The dataset provides natural-like yet controlled emotion data, which helps to develop reliable spontaneous emotion classifiers. The dataset was recorded at a very high frame rate capturing 120 frames per second, which enables the extraction of very fine facial changes that might help to analyse the emotion dynamics through time including the onset, apex, offset of facial expressions and the related emotion intensity.

Another advantage of using this dataset is the availability of contextual information, which is not available in most of the existing datasets using isolated sentences or short dialogues. The average duration of a conversation between two actors is approximately 5 minutes, and the sequential information may be used to improve the continuous emotion recognition. The categorical as well as continuous attribute-based annotation provides a detailed description of emotional content along with its intensity information.

4.6 Description of IEMOCAP

The IEMOCAP dataset was collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of South California (USC). In this dataset, ten skilled actors (5 males and 5 females) were recorded in 5 dyadic sessions during both scripted and improvised/spontaneous spoken communication scenarios. One of the two actors had a set of reflective markers on their face, head, and hands. The layout of the 53 facial markers, 2 hand markers, 4 wristbands markers, and 2 headbands markers is shown in Fig. 4.3. The facial markers provide detailed motion capture information about their facial expressions. High speed cameras capturing 120 frames

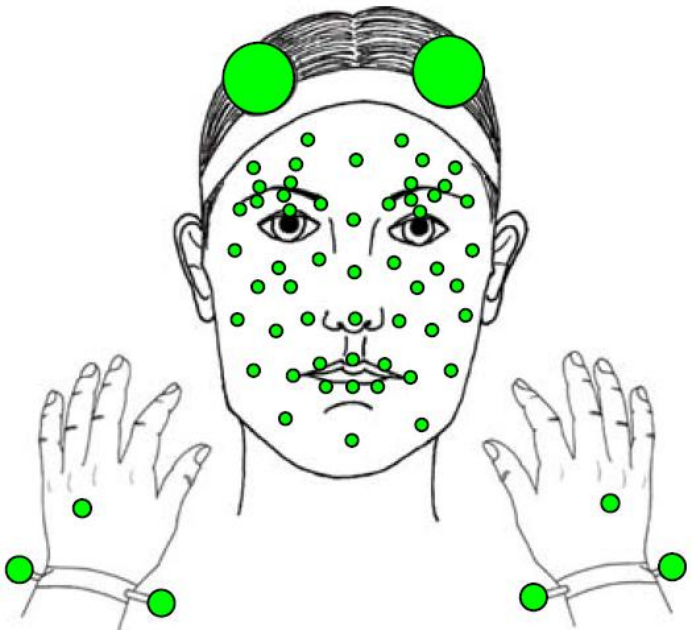


Figure 4.3: Marker Layout used in recording for IEMOCAP dataset [25].

per second were used to record the actors and the 3D positions of the facial markers was tracked with very high accuracy.

The released version of the dataset contains sets of points for fourteen improvised and fourteen scripted conversations lasting approximately 5 minutes each between two actors, one male and one female. For each conversation, only one of the two actors was recorded. Each conversation consists of almost 50 utterances of the two actors, where an utterance is a sentence or similar period during which one actor talks continuously. Each utterance consists of almost 11.4 words with an average duration of 4.5 seconds.

Within each conversation, each utterance of the two actors was annotated by three independent human evaluators into categorical labels (neutral, happiness, anger, sadness, surprise, disgust, fear, frustration, and excitement) as well as psychological data about emotion intensity (valence, activation, and dominance). These two approaches cover the categorical as well as continuous attributes of emotional representation.

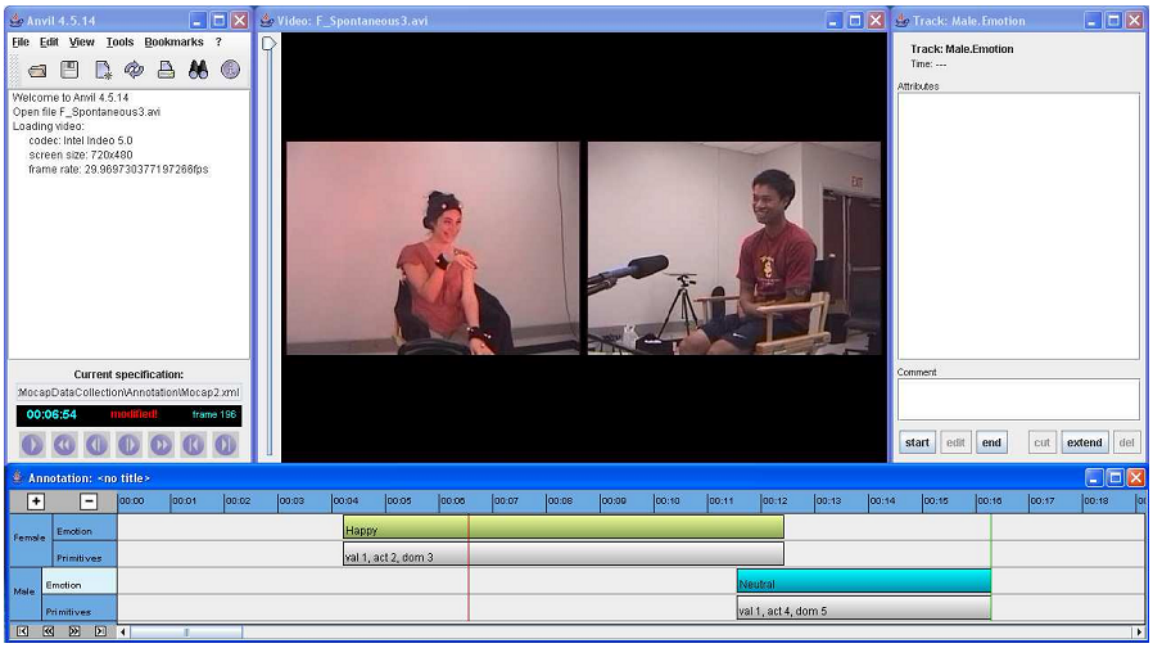


Figure 4.4: The ANVIL annotation tool used for emotion evaluation of the utterances based on categorical and continuous attributes [25].

The emotion content of the dataset was annotated by the evaluators into categorical labels using the ANVIL tool (see Fig. 4.4), after sequentially watching the videos. In this way, the evaluators had information from audio and video channels as well as previous utterances in the conversation for assessing the emotional content of each utterance. It was assumed that within an utterance, there is no emotion transition (e.g., from happy to excited), however, the evaluators were allowed to select more than one emotion category to describe mixtures/blends of emotions, which are more common in natural communication. The estimated confusion matrix between the assigned categorical labels shows that there is an overlap between happiness and excitement as well as anger and frustration. Neutral, disgust, and anger are often get confused with frustration. Also, sadness is often confused with frustration and neutral.

To evaluate continuous emotion attributes, SAMs were used. Two different human evaluators were asked to select the most suitable manikan (range: 1 to 5) for each

utterance. For valence, 1 means 'very negative' and 5 means 'very positive', similarly for activation, 1 means 'very passive' and 5 means 'very active'. The inter-evaluator agreement was higher for valence than activation and dominance. The values assigned by the evaluators show that most of the emotions in the dataset lie in the range of low to moderate intensity. For further details on any part of the data capture and labelling, see [25].

The selected dataset has some limitations such as, the released version contains motion capture data of only two subjects (one male and one female) recorded in fairly large sessions. The size of the dataset is quite large in terms of recorded data, but small in terms of participants. The evaluation-level limitation of the selected dataset is associated with its utterance-based annotation technique. Assuming the emotional content did not change much within an utterance (duration $\approx$ 4.5 seconds), the same label (or mixture of labels) was assigned to each frame in that utterance. The 'silent' frames and those containing sounds of active listening like 'mmhh' were not annotated at all. Another problem of the chosen dataset is that the human evaluators were sometimes inconsistent in labelling the data, since each one of them perceived and evaluated the emotions associated to an utterance in his own way. Due to the subjective nature of emotions, the inter-evaluator inconsistency is a common problem of all emotion datasets.

4.7 Data Preprocessing

We used the locations of the marker points in 3D as the basis of our analysis, and randomly chose 4,000 frames of each of six emotions (neutral, happiness, excitement, anger, frustration, and sadness) to form a training set of 24,000 frames, creating a separate set for each of the two actors. It should be noted that the marker points were

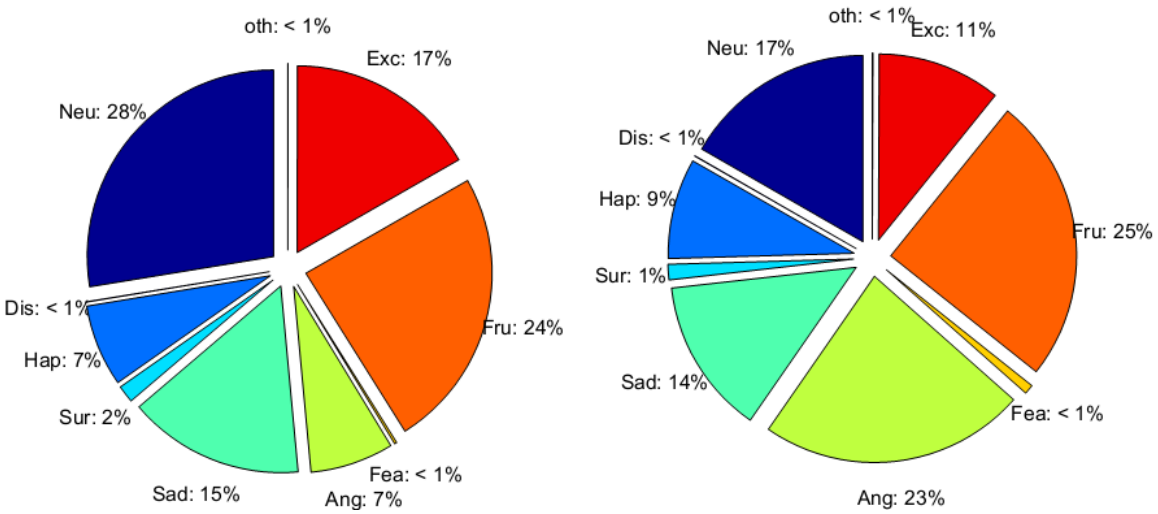


Figure 4.5: The distribution of data for each emotion category, *Neu*: Neutral, *Dis*: Disgust, *Hap*: Happiness, *Sur*: Surpsie, *Sad*: Sadness, *Ang*: Anger, *Fea*: Fear, *Fru*: Frustration, *Exc*: Excitement, *oth*: Other [25].

already aligned to make the nose marker at the center of each frame that removed any translation effects. The rotational effects were compensated by multiplying each frame by a rotational matrix. For details about markers alignment, refer to [25].

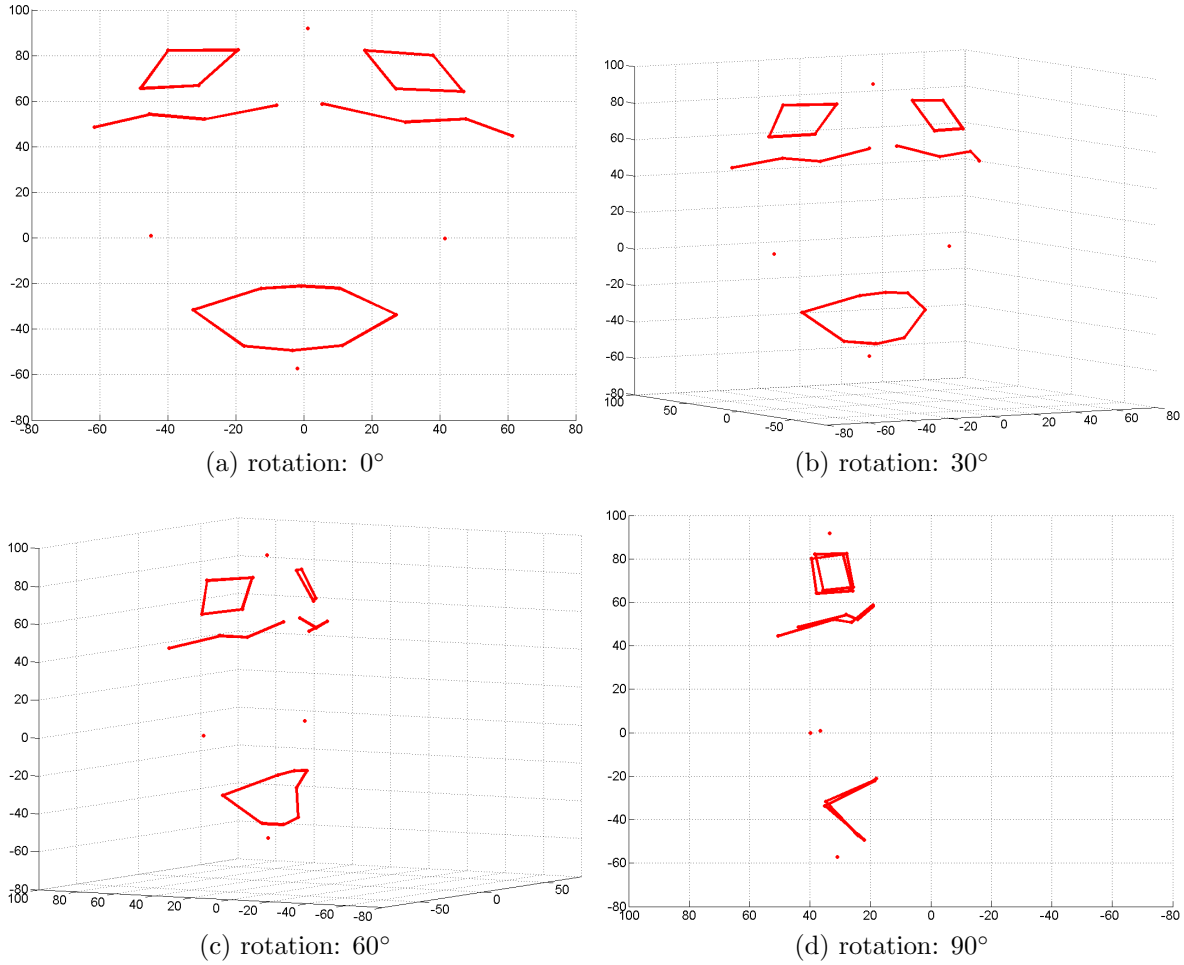
Each utterance was labelled by the three expert human evaluators in terms of discrete categories as well as emotion attributes (valence and activation). For the training set, we took frames from the utterances where all three experts agreed. We used six emotions rather than the full nine as for the missing emotions (disgust, surprise, and fear) there was insufficient data (as shown in Fig. 4.5), sometimes as little as 2,000 frames in total. Out of the six selected emotions, two (frustration and excitement) are the candidate basic emotions [68, 163]. We also selected seven continuous conversations comprising of almost 152,000 frames in total to form a testing set. For the testing set, there was no such condition of agreement by all three experts while choosing the frames.

Each frame of the dataset contains the motion capture information of 61 markers in 3 dimensions, so the training data was of size $24,000 \times 183$ dimensions. We reduced the dimensionality of the data for each frame in three ways:

1. Markers not on the face (such as the head and hands) were excluded.

2. Markers that did not move significantly (such as eyelids and nose) were removed.

3. Sets of markers that moved together (such as, points on the chin and forehead) were replaced by a single point at the centre of the set.



(a) rotation: 0°

(b) rotation: 30°

(c) rotation: 60°

(d) rotation: 90°

Figure 4.6: The 28 marker points in 3D used for emotion recognition and analysis.

As a result of these simplifications each emotion frame is represented by 28 markers points covering the forehead, eyebrows, eyes, cheeks, lips, and chin (Fig. 4.6). The location of marker points are in 3D, making it an 84D vector.

Out of the three attribute-based labels (valence, activation, and dominance), we have used just valence and activation values for two reasons: *first*, we chose to use the activation-evaluation space which defines emotions in terms of these two dimensions, *second*, there is a considerable disagreement in defining the third dimension (dominance) for emotion representation in the psychological literature.

Due to the confusion between neutral and sadness, mostly the neutral state was assigned an attribute value of 3 for valence and 1 (very passive) for activation. To correct this, based on the majority voting of categorical labels, we change the activation value back to 3 in the case of the neutral state. This correction is also supported by the model of activation-evaluation space, where neutral lies at the centre. Following the assumption that neutral state is a 'no emotion' state, its position in the space does not effect the position of other five emotions.

The segmentation of a conversation into utterances is useful for discrete emotion recognition, however, it is a problem for continuous emotion recognition and analysis. During a conversation, there are several places where both actors were silent, which caused an interruption in the continuous information. For this reason, we asked Dr. Carlos Busso (Assistant Professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD), who is one of the authors of the IEMOCAP dataset) to provide us with the unsegmented continuous data for uninterrupted analysis of emotions within full conversations. The continuous conversations include the frames where both actors were silent as well as the overlapping frames where both of them were speaking at the same time. The overlap usually appears at the end of first actor's utterance where the second actor starts talking. This type of overlap is

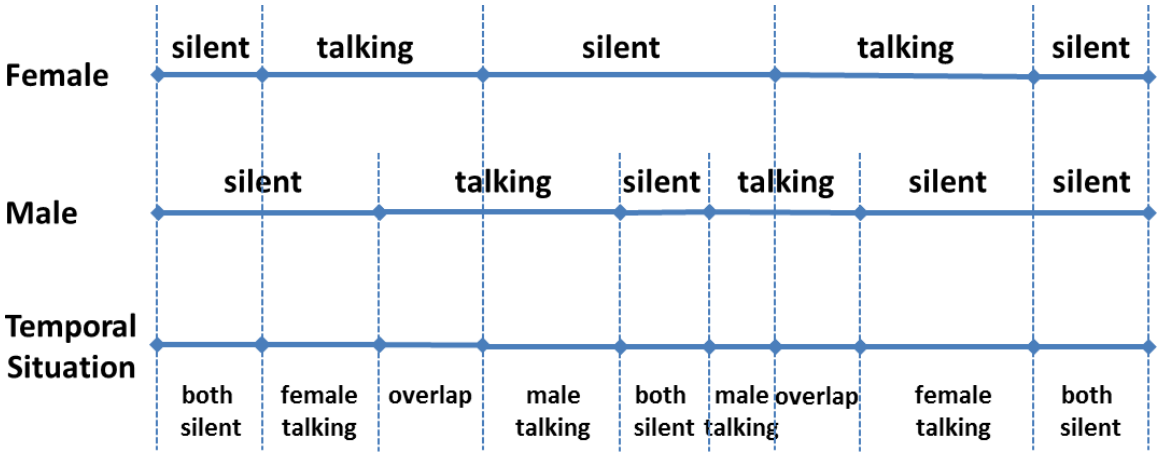


Figure 4.7: An example of data layout in a continuous conversation between two actors (male and female). The term 'overlap' refers to the situation where both actors are talking at the same time.

quite common in natural communication. An example of the structure of continuous data in a conversation between two actors (one male and one female) is shown in the Fig. 4.7.

4.8 Summary

This chapter has presented the basic criteria for the selection of an appropriate dataset for the task of emotion recognition and analysis. A review of the most commonly-used video and audiovisual datasets was presented with respect to the given criteria. The chapter also demonstrated a review of existing tools used for annotating emotion data. Based on the critical review of the five datasets fulfilling most of the given criteria, we chose to select the IEMOCAP dataset for emotion modelling and analysis. The marker layout, segmentation process, and annotation details of the selected dataset was also demonstrated. Finally, the necessary data preprocessing steps were discussed.

Chapter 5

Basic Emotion Recognition

5.1 Introduction

Following the psychological assumption that whenever we feel an emotion, it appears on our face, we started with the classification of facial changes caused by expressions into the basic emotion categories. In this chapter, we will show that building statistical shape models of different parts of the face and combining them can give more successful results than using only a model of the whole face. The statistical shape modelling based on Principal Component Analysis is first described, and then the classification technique is discussed. The detailed analysis of the shape models is presented, together with experiments showing the effectiveness of the proposed method for recognising discrete basic emotions.

5.2 Shape Models

As described in Chapter 4, we are using the 3D locations of 28 marker points on the face for emotion recognition and analysis. Given that, as a result of this, we have 84D

datapoints containing the motion capture information of facial points, the question is how to recognise emotions.

The models used for the analysis of shape variation marked with landmark points are referred to as statistical shape models, also known as point distribution models (PDM) [38, 39]. A usual implementation of these models is to use them as Active Shape Models (ASM) [40]. ASMs are also known as 'Smart Snakes' as they both are deformable models but contrary to Snakes [121] , ASMs have global constraints with respect to shape which they learn from observations in the set of labelled examples. ASM is a method of fitting the model points to the new image points by iteratively changing the model points to get the best possible fit. We have used shape models to analyse variations in the 84D datapoints and classified the unknwon set of points based on the minimum distance between the set of points and the mean of each shape model separately. Based on the assumption that the location of marker points in the training set is accurate, we do not need an iterative refinement process to improve the distance between the model points and unknown points.

The first step is to find a simplified representation of such high-dimensional data in order to visualise and understand the relationships among multiple variables. Generally, in a multivariate dataset there is more than one variable measuring the same kind of behaviour. The problem may be simplified by replacing such redundant groups of variables by a single new variable.

A standard technique to model shape variation and analyse sets of datapoints in high dimensional spaces is Principal Component Analysis (PCA) [115, 141]. PCA finds a new set of variables, called *principal components* (PCs), by identifying a linear transformation (translation, rotation, and scaling) of the original variables in the dataset. All principal components are mutually orthogonal, such that ideally there is no redundant information. In this case, no redundancy means that the

principal components are uncorrelated with each other. Each component accounts for a maximal amount of variance in the observed variables that was not accounted for by the preceding components, and is therefore uncorrelated with all of the preceding components. The principal components are statistically independent to each other only for normal (Gaussian) random variables [115]. As a whole, the set of principal components form an orthogonal basis for the space of the original dataset. The resultant basis has maximum variance of the dataset along the first basis vector, and successively less variance amongst the following basis vectors.

To define principal component analysis and explain the computation of principal components, we follow the notation used by Jolliffe in [115]. Given a vector $\mathbf{x}$ of p random variables, we are interested to find a linear function $\mathbf{a_1^T x}$ of the elements of $\mathbf{x}$ having maximum variance. $\mathbf{a_1}$ is a vector of length p such that the sum of square of elements of $\mathbf{a_1}$ equals 1, and the superscript $\mathbf{T}$ represents the transpose,

$$\mathbf{a_1^T x} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = \sum_{j=1}^{p} a_{1j}x_j \tag{5.1}$$

Next we are looking for a linear function $\mathbf{a_2^T x}$, uncorrelated with $\mathbf{a_1^T x}$ having maximum variance, and so on. At stage k, a linear function $\mathbf{a_k^T x}$ which is uncorrelated with all previous functions $\mathbf{a_1^T x}, \mathbf{a_2^T x}, \cdots, \mathbf{a_{k-1}^T x}$ is found with the maximum variance. In this way, $\mathbf{a_1^T x}$ is the first PC having the maximum variance of the data, $\mathbf{a_2^T x}$ is the second PC, having less variance than the first PC, and successively, $\mathbf{a_k^T x}$ is the k^{th} PC, which has less variance than the previous $(k-1)$ PCs. For optimal low dimensional approximation, $\mathbf{x}$ should be mean-centered (mean-subtracted). This is to make sure that the first principal component refers to the direction of maximum variance of the data instead of corresponding to the mean of the data [153]. Up to p PCs could be

found, but generally, it is hoped that the first m PCs ($m << p$) adequately account for the majority of the variance of data.

Let Σ represents the covariance matrix of $\mathbf{x}$ whose $(i,j)th$ element represents the covariance between the ith and jth elements of $\mathbf{x}$ when $i \neq j$, and the variance of the jth element when $i = j$. It turns out that for $k = 1, 2, \cdots, p$, $\mathbf{a_k^T x}$ is the kth PC where $\mathbf{a_k}$ is an eigenvector of Σ corresponding to its kth largest eigenvalue λ_k. λ_k denotes the variance of the kth PC. We applied PCA to each mean-centered 84D datapoint using the Matlab's *princomp* from the Statistics toolbox.

In many applications using PCA, the objective is to reduce the number of PCs (m) to a smaller number without losing much information. There are several ways to choose a suitable subset of principal components; we have used the *scree graph*. Since, the variance of the k^{th} principal component is λ_k, the scree graph is a plot of λ_k against k. Fig. 5.1 shows the scree graph for the full face data plotting the percentage variance of the first ten principal components. We selected the first five principal components, which covered 93% of the total variation of the data. The selection of first few principal components reduces the dimensionality of data to a great extent (i.e., from 24,000 $\times$ 84 dimensions to 24,000 $\times$ 5 dimensions), while retaining as much as possible variation of the dataset.

We then examined the effect of each PC independently, by applied changes of ± 3 standard deviations along each PC. We noticed that the first PC (PC1) of the full face, which covers almost 50% of the total variation, was describing the upward and downward movement of the mouth points (i.e., lips). Fig. 5.2 shows the effects of varying PC1 ($\mu \pm 3\sigma$, where μ denotes the mean face and σ denotes the standard deviation) for the female actor. We hypothesised that this was due to talking, and therefore selected a set of silent frames from the data and applied PCA to these points. Since the first PC was not present in the silent frames, we decided that it was

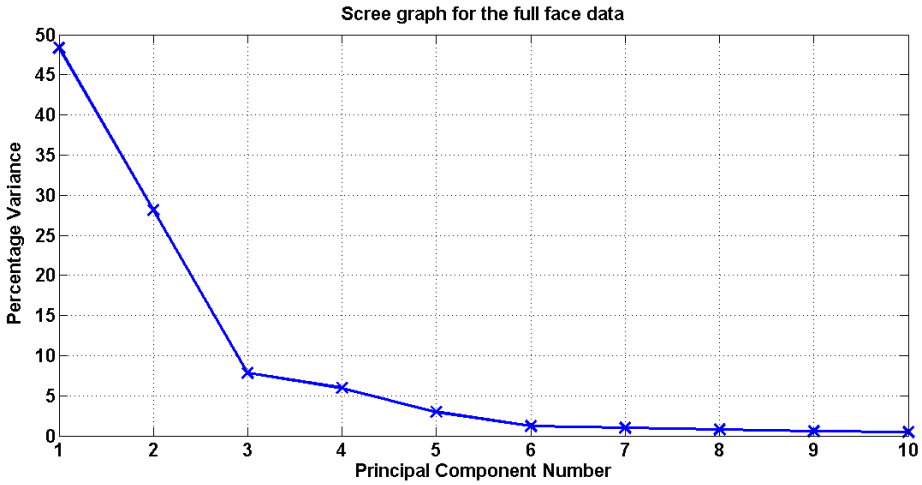


Figure 5.1: The scree graph for the full face data plotting percentage variance covered by the first ten principal components.

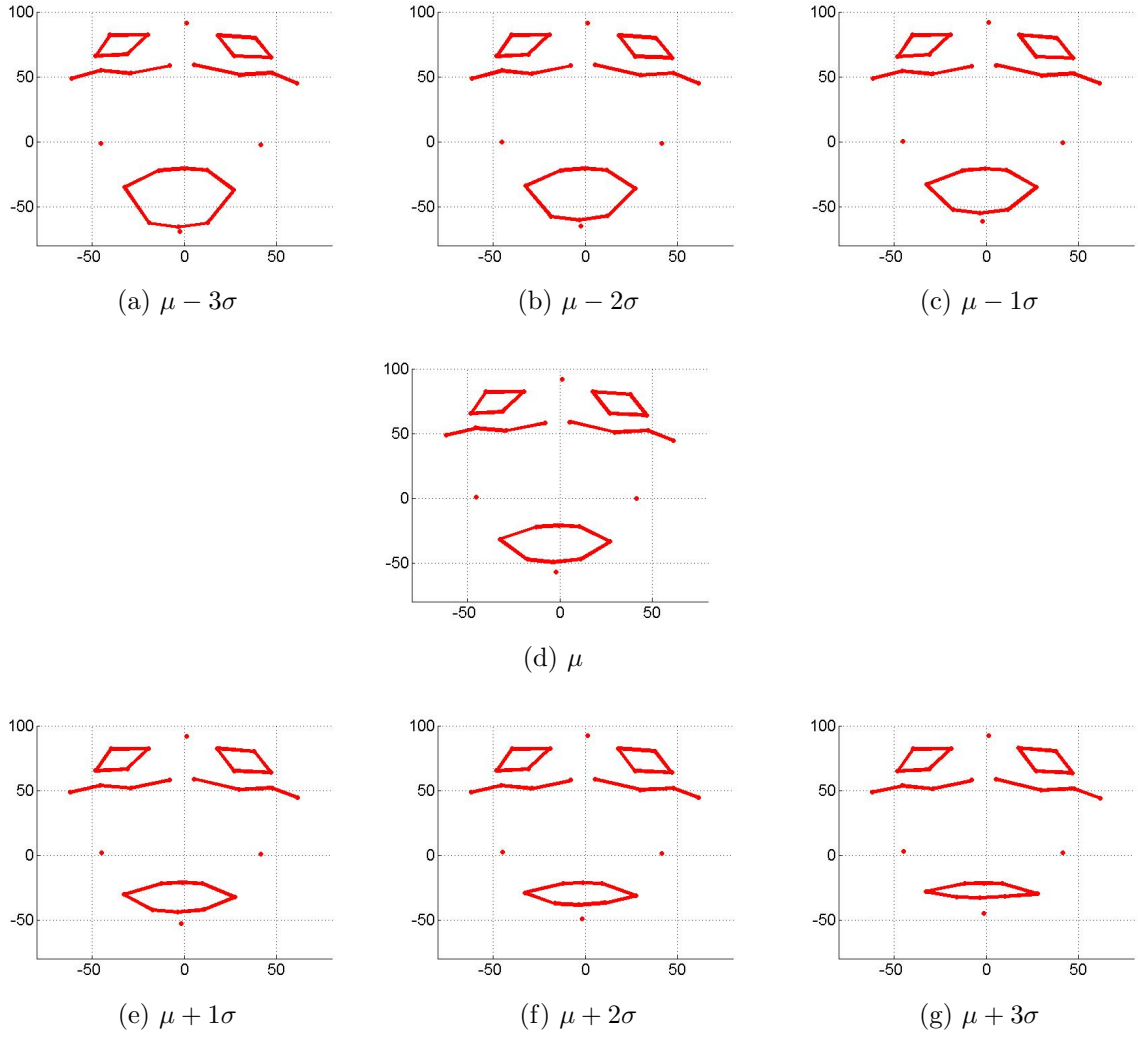
primarily correlated with talking and therefore not directly connected with emotion recognition, and so we discarded it. In fact, for emotion recognition, talking is often found to be one of the biggest confusion factors and sources of error [199].

5.2.1 Using the upper and lower face separately

According to [13], some expressions are better recognized from muscle activity in the upper half of the face, while others use muscles primarily from the lower half of the face. Moreover, [74] suggested that the upper face (including the eyebrows and forehead) is difficult to control voluntarily as compared to the lower face including the mouth, cheeks and chin. On the basis of these observations, we also created separate shape models of the upper and lower halves of the face. Within the IEMOCAP dataset, the upper-half contains 17 marker points covering the forehead, eyebrows, and eyes as shown in Fig. 5.3; while the lower-half consists of 11 marker points covering the cheeks, lips, and chin as shown in Fig. 5.4. Fig. 5.5(a) plots the scree graph of the upper face data, and Fig. 5.5(b) plots the scree graph of the lower face data. For the upper face, we selected the first 4 PCs, which covered 93.4% of the total variance

of the data. For the lower face, 4 PCs (2-5) were selected, covering almost 95% of the total variance of the data. The first PC of the lower face was found to be correlated with talking, and so was discarded.

Consequently, we chose to use four PCs (2-5) of full and lower face model and four PCs (1-4) of the upper face model for our analysis. We transformed the training data into the three different 4D spaces of these sets of four principal components. Each datapoint was then labelled with the majority vote of the three human experts, so



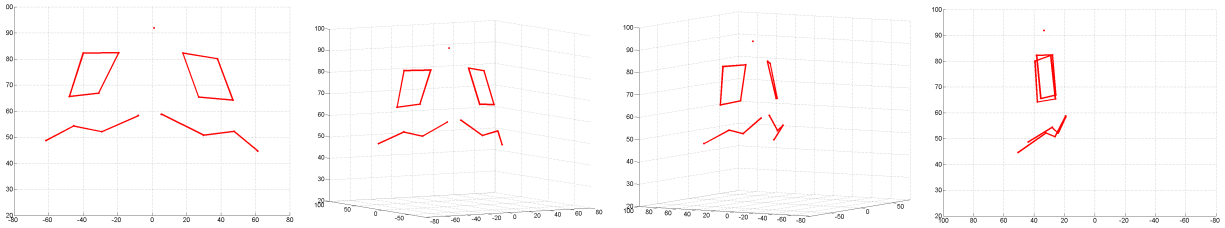
(a) $\mu - 3\sigma$ (b) $\mu - 2\sigma$ (c) $\mu - 1\sigma$

(d) μ

(e) $\mu + 1\sigma$ (f) $\mu + 2\sigma$ (g) $\mu + 3\sigma$

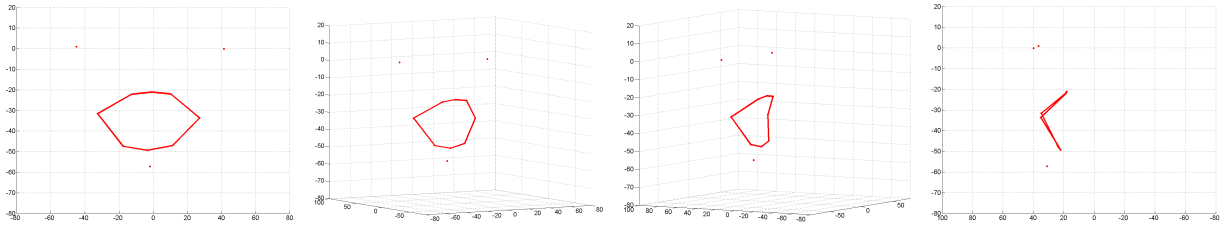
Figure 5.2: The effect of varying the first PC (PC1) for the full face of female actor. μ denotes the mean face and σ denotes the standard deviation.

that the training set consists of 24,000 points, each labelled as one of six emotions in three different 4D spaces.



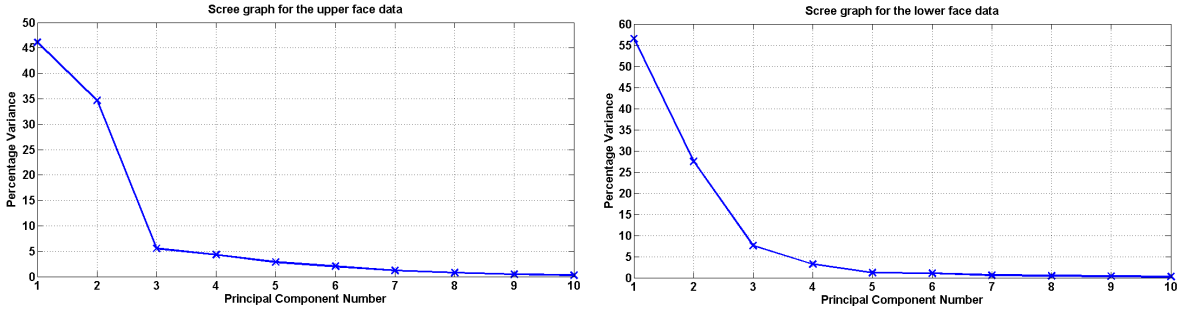
(a) rotation: 0° (b) rotation: 30° (c) rotation: 60° (d) rotation: 90°

Figure 5.3: The 17 marker points in 3D on the mean upper face of the female actor.



(a) rotation: 0° (b) rotation: 30° (c) rotation: 60° (d) rotation: 90°

Figure 5.4: The 11 marker points in 3D on the mean lower face of the female actor.



(a) (b)

Figure 5.5: The scree graphs plotting the percentage variance covered by first ten principal components for (a) the upper face data, and (b) the lower face data.

5.2.2 Classification

For classification, we replaced each cluster (i.e., set of points labelled as one emotion) with the mean of that set, and also computed the covariance matrix (spread) of the data. We therefore ended up with six datapoints representing the mean of each set and an associated covariance matrix.

For classification of a test frame, it was transformed into the 4D space of each model separately. We then computed the Mahalanobis distance between the test frame and each of the six emotion clusters. In this way, we got three sets of six distances; one for each emotion in each model space (total 18 distances). For each emotion we take the minimum distance across the three models. Fig. 5.6 shows the block diagram of the proposed joint face model. We then labelled the test point with the label of the cluster that it is closest to. The Mahalanobis distance not only uses the mean, but also takes into account the spread of the data to compute a distance. It is formulated as:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

where $\mathbf{x}$ is the (4D) column vector of the test frame, μ is a column vector of the mean and Σ is the 4×4 covariance matrix for an emotion. If the covariance matrix is set to the identity matrix, then the Mahalanobis distance reduces to the Euclidean distance [137, 141].

Computing the Mahalanobis distance is a computationally expensive process which calculates the covariance matrix and then its inverse. For efficient computation, we have used Matlab's *mahal* from the Statistics toolbox to compute the Mahalanobis distance of each frame to each of the six emotion clusters.

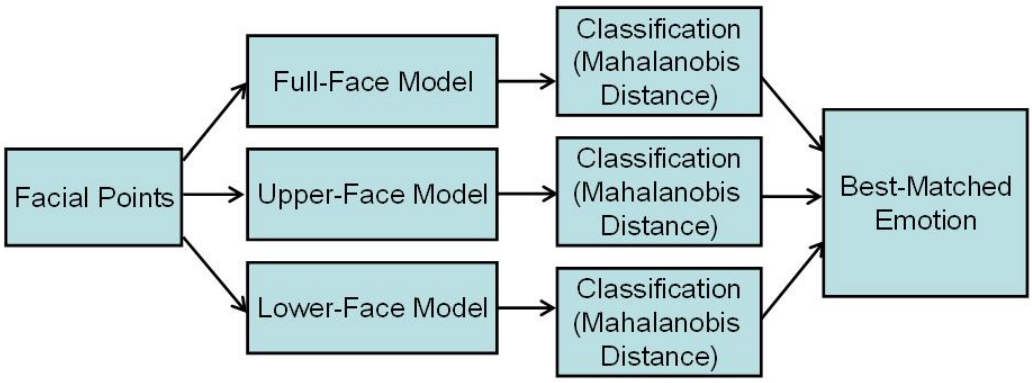


Figure 5.6: The proposed joint face model.

5.3 The Effects of Each Principal Component

Having already identified that PC1 was involved in talking, we chose to investigate the effects of the other four PCs of the full model as well. PC2 covers 28% of the total variation of data. It controls the sideways movement of mouth points (lips). The outward movement of the lips contribute to the positive expressions (e.g., smile) while the inward movement gives the negative expressions (e.g., sad). The third PC covers 7.8% of the total data variation and contributes to the upward and downward movement of eyebrows and forehead marker points, while PC4 covers 6% of the total variation and contributed to the sideways movement (i.e., towards and away from the nose) of eyebrows and forehead marker points. The fifth PC covers 3% of the variation of data and appears as a rather strange circular motion of the lips. Experimentally we observed that this PC contributed mainly to the laughing expression.

Table 5.1 describes the effect of each PC on the mean face, which is the same for both the female and male full face shape models. In the table, '+' means moving the mean face μ for positive weights λ, formulated as: $\mu + \lambda PC_i$ (where μ is the transformed mean face and PC_i is the ith principal component, while '-' means using negative weights λ). Fig. 5.7 shows the effects of varying PC2 ($\mu \pm 3\sigma$, where μ

$PC2$	$+$	Outward movement of lips
$PC2$	$-$	Inward movement of lips
$PC3$	$+$	Upward movement of eyebrows and forehead
$PC3$	$-$	Downward movement of eyebrows and forehead
$PC4$	$+$	Inward movement of eyebrows and forehead
$PC4$	$-$	Outward movement of eyebrows and forehead
$PC5$	$+$	Right to left circular movement of lips
$PC5$	$-$	Left to right circular movement of lips

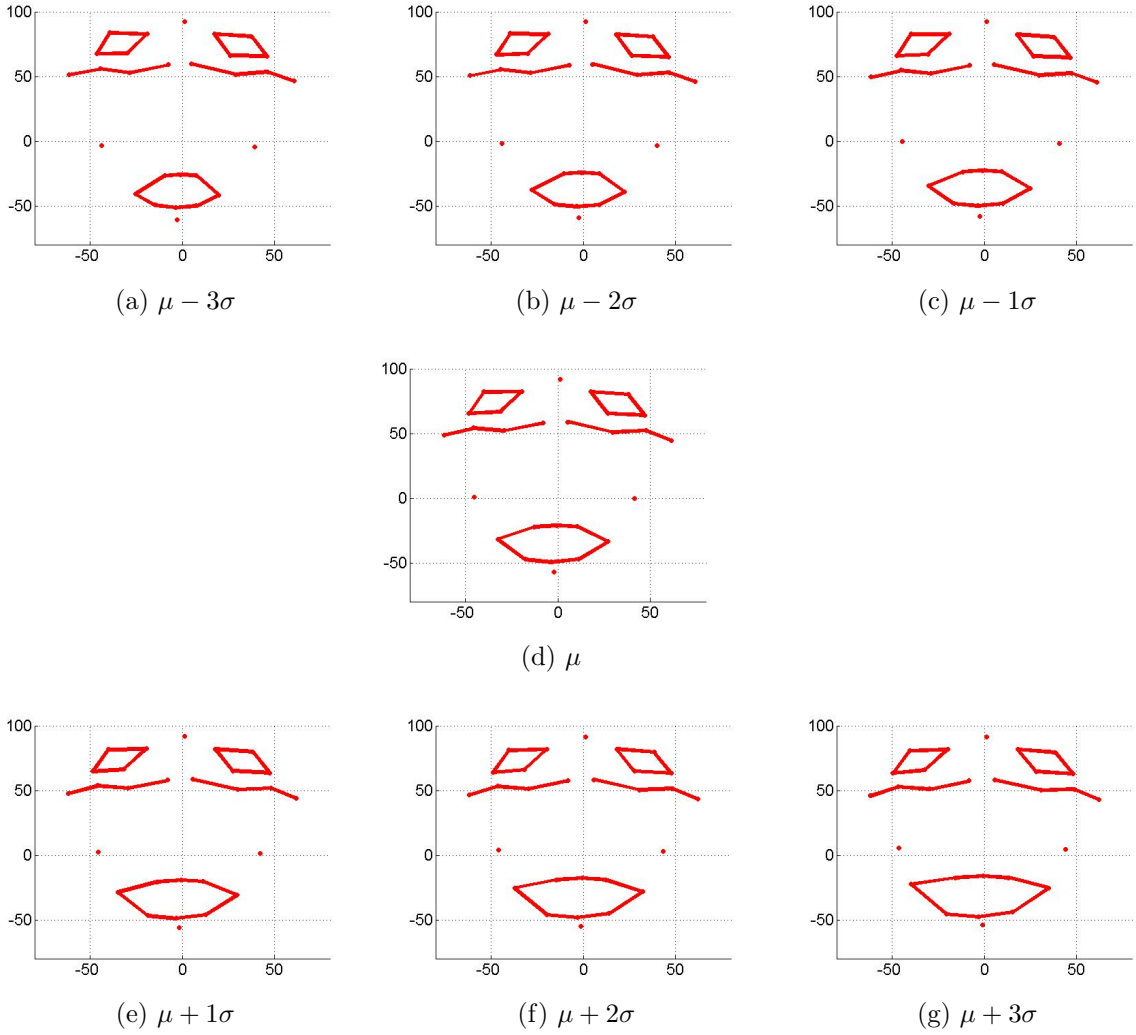
Table 5.1: The effect of each PC on the mean face. For explanation, see the text.

denotes the mean face and σ denotes the standard deviation) for the full face of the female actor.

5.3.1 Effect of each full, upper, and lower face PC

For the *upper-face*, the first PC covers almost 46% variation. Since this model cannot see the effects of talking, this PC identifies upward motion of the eyebrows and forehead, while the second PC also deals with similar movement and covers 34.6% of the variation. Fig. 5.8 shows the effects of varying PC2 ($\mu \pm 3\sigma$, where μ denotes the mean face and σ denotes the standard deviation) for the upper face of the female actor. Table C.1 in Appendix C listed the numerical values demonstrating the direction of 3D movement of 17 marker points between $\pm 3\sigma$. Considering the mean face as a reference, the values identify the upward movement of forehead and eyebrows marker points by varying PC2 from -1σ to -3σ and the downward movement of forehead and eyebrows marker points by varying PC2 from $+1\sigma$ to $+3\sigma$. The third PC, which covers 5.5% of the variation highlights the inner brows moving upward, and the fourth PC (4.3%) identifies the outer brow moving in the upward direction.

In the case of the *lower-face* model, the first PC covers 56.5% of the variation of data and identifies the upward and downward movement of the mouth points. This movement of mouth points was experimentally shown to be highly correlated with

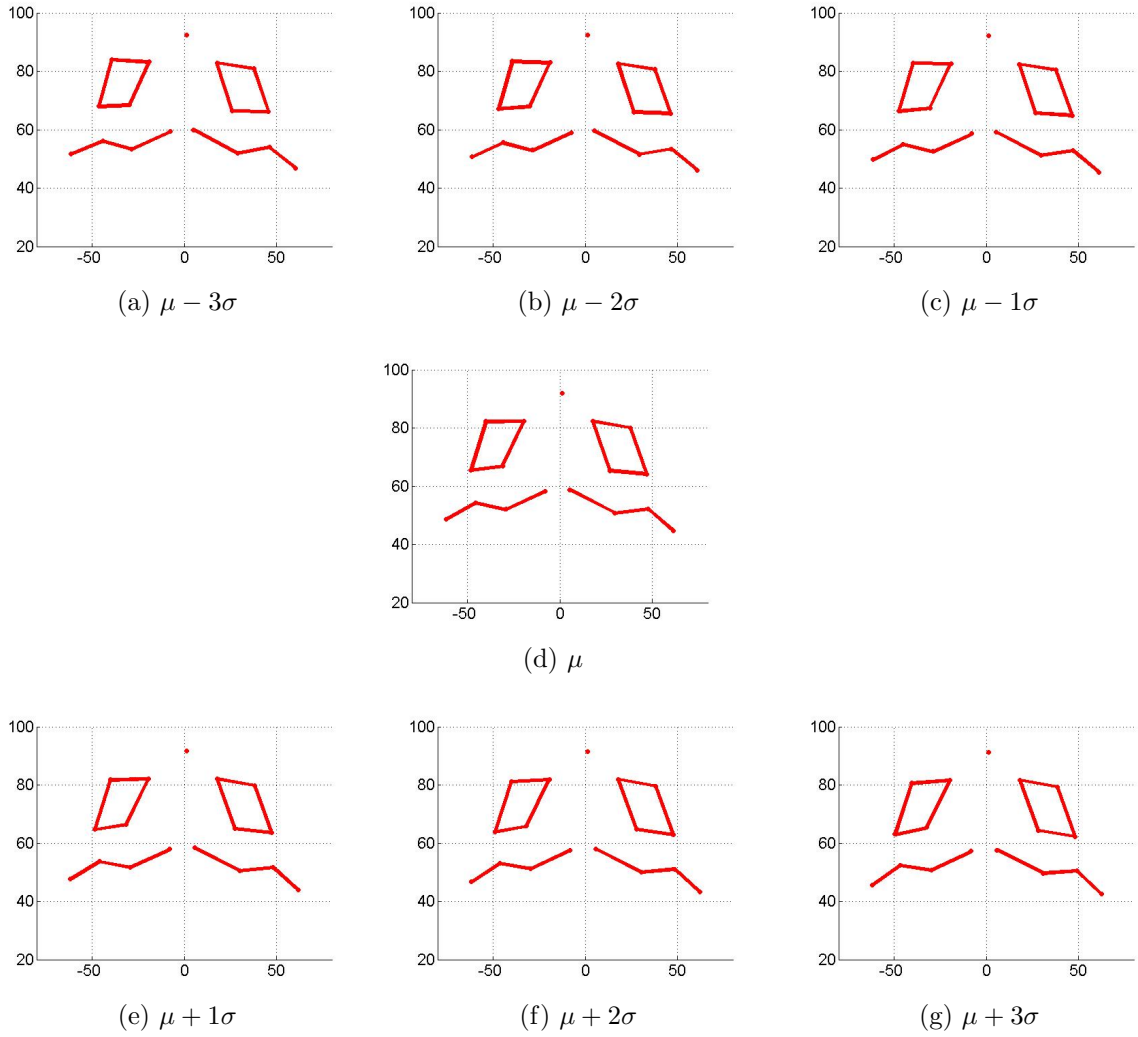


(a) $\mu - 3\sigma$ (b) $\mu - 2\sigma$ (c) $\mu - 1\sigma$

(d) μ

(e) $\mu + 1\sigma$ (f) $\mu + 2\sigma$ (g) $\mu + 3\sigma$

Figure 5.7: The effects of varying the second principal component (PC2) on the female full face model. μ denotes the mean face and σ denotes the standard deviation.



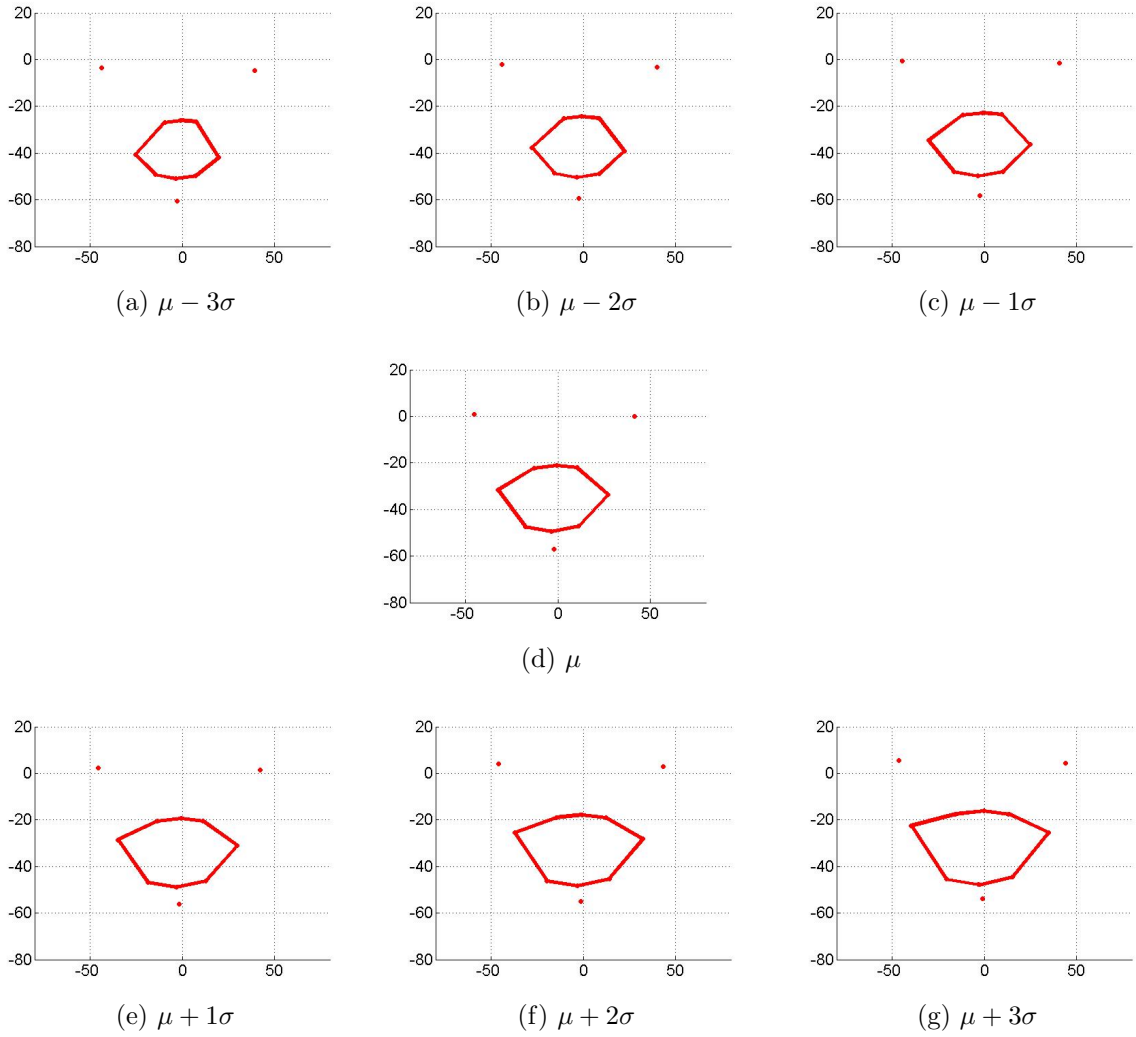
(a) $\mu - 3\sigma$ (b) $\mu - 2\sigma$ (c) $\mu - 1\sigma$

(d) μ

(e) $\mu + 1\sigma$ (f) $\mu + 2\sigma$ (g) $\mu + 3\sigma$

Figure 5.8: The effects of varying the second principal component (PC2) on the female upper face. μ denotes the mean face and σ denotes the standard deviation. The numerical values demonstrating the direction of variation of marker points are listed in Table C.1 in Appendix C.

talking, and so was discarded. The second PC covers 27.5% variation deals with the sideways movement of mouth. Fig. 5.9 shows the effects of varying PC2 ($\mu \pm 3\sigma$) on the lower face of female actor. The first and second PCs of the lower face are same as that of the full face PCs. The third PC covers 7.6% of the variation and highlights the upward and downward movement of cheeks, the fourth PC covers 3.2% of the variation identifies the upward and downward movement of chin, while the fifth PC



(a) $\mu - 3\sigma$ (b) $\mu - 2\sigma$ (c) $\mu - 1\sigma$

(d) μ

(e) $\mu + 1\sigma$ (f) $\mu + 2\sigma$ (g) $\mu + 3\sigma$

Figure 5.9: The effects of varying the second principal component (PC2) on the female lower face. μ denotes the mean face and σ denotes the standard deviation.

covers 2% of the variation identifies the circular movement of lips, similar to the fifth PC of the full face.

It should be noted that all the figures are actually in 3D, but to get a clear view the viewpoint has been rotated by setting the azimuth and elevation equal to zero.

5.4 Performance Comparison

Based on the shape models we had four methods of classifying emotions: using the joint model, or any one of the three models separately. As a comparison, we implemented two methods that have been used in the past: a set of rule-based classifiers [107] and a set of Support Vector Machines (SVM) [151]. These methods are described in this section.

5.4.1 A rule-based emotion classifier

In the previous section we looked at the effect of each principal component separately. This enabled us to create a simple rule-based classifier by observing how the PCs moved and effected the mean face. The rules we applied are listed in the second column of Table 5.2.

The rules presented in the second column of Table 5.2 also highlight something that is known from psychology, and that we observed in our data, namely that anger and frustration look similar on the face, as do happiness and excitement. These clusters of points overlap significantly, as shown in Fig. 5.10(a) and Fig. 5.10(b). According to these rules, anger and frustration share the negative value of PC2 and PC4 and happiness and excitement both have the positive value of PC2. Happiness and anger are well separated (see Fig. 5.10(c)), while sadness and anger also overlap at some points (see Fig. 5.10(d)) since both have the negative value of PC2.

	Rules based on PCs	**MPEG-4 FAPs**
Happiness	$PC2$ is positive, $PC3$ is positive, $PC2$ has larger magnitude than $PC3$	open_jaw (F_3), lower_t_midlip (F_4), raise_b_midlip (F_5), stretch_l_cornerlip (F_6), stretch_r_cornerlip (F_7), raise_l_cornerlip (F_{12}), raise_r_cornerlip (F_{13}), close_t_l_eyelid (F_{19}), close_t_r_eyelid (F_{20}), close_b_l_eyelid (F_{21}), close_b_r_eyelid (F_{22}), raise_l_m_eyebrow (F_{33}), raise_r_m_eyebrow (F_{34}), lift_l_cheek (F_{41}), lift_r_cheek (F_{42}), stretch_l_cornerlip_o (F_{53}), stretch_r_cornerlip_o (F_{54})
Excitement	$PC2$ is positive, $PC3$ is positive, $PC3$ has larger magnitude than $PC2$	N/A
Anger	$PC2$ is negative, $PC3$ is positive, $PC4$ is negative	lower_t_midlip (F_4), raise_b_midlip (F_5), push_b_lip (F_{16}), depress_chin (F_{18}), close_t_l_eyelid (F_{19}), close_t_r_eyelid (F_{20}), close_b_l_eyelid (F_{21}), close_b_r_eyelid (F_{22}), raise_l_i_eyebrow (F_{31}), raise_r_i_eyebrow (F_{32}), raise_l_m_eyebrow (F_{33}), raise_r_m_eyebrow (F_{34}), raise_l_o_eyebrow (F_{35}), raise_r_o_eyebrow (F_{36}), squeeze_l_eyebrow (F_{37}), squeeze_r_eyebrow (F_{38})
Frustration	$PC2$ is negative, $PC3$ is negative, $PC4$ is negative	N/A
Sadness	$PC2$ is negative, $PC3$ is negative, $PC4$ is positive	close_t_l_eyelid (F_{19}), close_t_r_eyelid (F_{20}), close_b_l_eyelid (F_{21}), close_b_r_eyelid (F_{22}), raise_l_i_eyebrow (F_{31}), raise_r_i_eyebrow (F_{32}), raise_l_m_eyebrow (F_{33}), raise_r_m_eyebrow (F_{34}), raise_l_o_eyebrow (F_{35}), raise_r_o_eyebrow (F_{36})

Table 5.2: Rules for classifying emotions based on the direction and magnitude of the selected full face principal components and the corresponding vocabulary of FAPs associated with the expression of each basic emotion. The vocabulary of FAPs is taken from [181]. The description of each of the mentioned FAP is listed in Table 5.3. The relationship between PCs and the corresponding FAPs is discussed in the text.

FAP Name	**FAP Number**	**Description**
open_jaw	F_3	Vertical jaw displacement
lower_t_midlip	F_4	Vertical top middle inner lip displacement
raise_b_midlip	F_5	Vertical bottom middle inner lip displacement
stretch_l_cornerlip	F_6	Horizontal displacement of left inner lip corner
stretch_r_cornerlip	F_7	Horizontal displacement of right inner lip corner
raise_l_cornerlip	F_{12}	Horizontal displacement of left inner lip corner
raise_r_cornerlip	F_{13}	Horizontal displacement of right inner lip corner
push_b_lip	F_{16}	Depth displacement of bottom middle lip
depress_chin	F_{18}	Upward and compressing movement of the chin
close_t_l_eyelid	F_{19}	Vertical displacement of top left eyelid
close_t_r_eyelid	F_{20}	Vertical displacement of top right eyelid
close_b_l_eyelid	F_{21}	Vertical displacement of bottom left eyelid
close_b_r_eyelid	F_{22}	Vertical displacement of bottom right eyelid
raise_l_i_eyebrow	F_{31}	Vertical displacement of left inner eyebrow
raise_r_i_eyebrow	F_{32}	Vertical displacement of right inner eyebrow
raise_l_m_eyebrow	F_{33}	Vertical displacement of left middle eyebrow
raise_r_m_eyebrow	F_{34}	Vertical displacement of right middle eyebrow
raise_l_o_eyebrow	F_{35}	Vertical displacement of left outer eyebrow
raise_r_o_eyebrow	F_{36}	Vertical displacement of right outer eyebrow
squeeze_l_eyebrow	F_{37}	Horizontal displacement of left eyebrow
squeeze_r_eyebrow	F_{38}	Horizontal displacement of right eyebrow
lift_l_cheek	F_{41}	Vertical displacement of left cheek
lift_r_cheek	F_{42}	Vertical displacement of right cheek
stretch_l_cornerlip_o	F_{53}	Horizontal displacement of left outer lip corner
stretch_r_cornerlip_o	F_{54}	Horizontal displacement of right outer lip corner

Table 5.3: Description of each MPEG-4 FAPS mentioned in Table 5.2.

We compared these rules with the vocabulary of facial animation parameters (FAPs) required for the description of facial expressions of basic emotions as presented in [181]. FAPs are a set of parameters defined in the MPEG-4 standard [165] for facial animation purposes. These parameters are strongly related to the action units (AUs) which describe the smallest visually perceptible facial movements. There are 66 low-level FAPs used to describe the movements of facial features including

eyes, nose, lips, jaw, cheeks, mouth, ears. Table 5.2 lists the vocabulary of FAPs associated with the expression of happiness, anger, and sadness. Since frustration and excitement are the candidate basic emotions we could not find the FAPs associated with the facial expressions of these emotions. The description of each of the mentioned FAP is listed in Table 5.3. The vocabulary for each emotion lists all FAPs that may be associated with the expression of that particular emotion; this does not mean that they all would necessarily produce the corresponding emotion [181].

For an expression of happy emotion, we observed that PC2 (which is associated with the horizontal movement of lips) varies in the positive direction. The horizontal movement of lips is described by the facial animation parameters F_6, F_7, F_{12}, F_{13} and the associated movement of cheeks is described by F_{41}, F_{42}, F_{53}, and F_{54}. We have excluded eyelids markers to avoid false poitives caused by blinking, so did not observe F_{19}, F_{20}, F_{21}, and F_{22} describing eyelids movements. F_{33} and F_{34} are associated with the vertical displacement of eyebrows, which is described by the variation of PC3 in positive direction. We did not observe F_3, F_4, and F_5 for an expression of happy emotion in our data. These FAPs are used to describe laughing (open mouth) expression of happiness, which in our case is described by PC5.

For angry emotion, we observed that PC2 varies in the negative direction. The horizontal inward movement of lips is described by F_{16} and the associated chin movement is described by F_{18}. We did not observe F_{19}, F_{20}, F_{21}, and F_{22} describing eyelids movements. F_{31}, F_{32}, F_{33}, F_{34}, F_{35}, and F_{36} are associated with the vertical displacement of eyebrows, which is described by the variation of PC3 in negative direction. F_{37} and F_{38} are associated with the horizontal displacement of eyebrows, which is described by the variation of PC4 in negative direction.

For an expression of sad emotion, we observed the variation of PC2 in negative direction but FAPs vocabulary does not include any such movement of lips. We

did not observe F_{19}, F_{20}, F_{21}, and F_{22} describing eyelids movements. F_{31}, F_{32}, F_{33}, F_{34}, F_{35}, and F_{36} are associated with the vertical displacement of eyebrows, which is described by the variation of PC3 in positive direction. We also observed the variation of PC4 associated with the horizontal inward displacement of eyebrows in sad expression, which is not included in FAPs vocabulary.

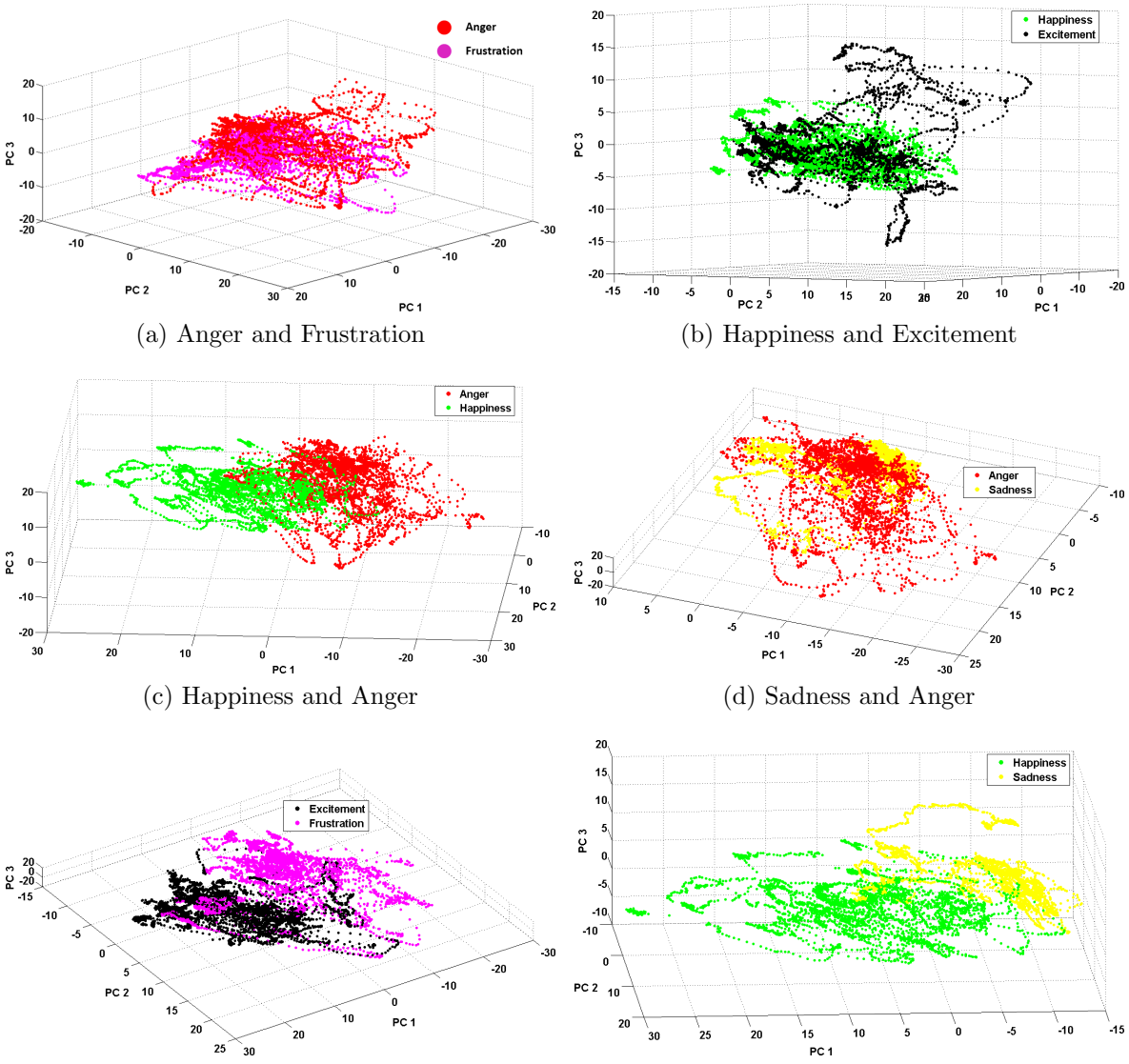
5.4.2 Support Vector Machines

The Support Vector Machine (SVM) is a popular machine learning technique for emotion recognition [151]. We used an SVM with a quadratic kernel using Sequential Minimal Optimization (SMO) method [172], implemented using Matlab's *svmclassify* from the Bioinformatics toolbox. The SVM is a binary classifier, so we convert it to a multi-class classifier by using one-versus-all training, where a separate SVM is trained to recognise each class, compared to all of the others. As a result, we obtained six SVMs, one for each emotion.

Each SVM is trained on the training data with the labels where at least two human observers agreed. It should be noted that the SVMs are not trained on the transformed data in 4D space using PCA. Classification was then performed on the basis of class membership (± 1), where $+1$ means successfully matched to the training class.

5.5 Results

In this section we report two experiments: first, comparing the four shape models, the rule-based classifier, and the SVM-based classifier based on training with $24,000$ datapoints in either six or four classes, and second, examining the robustness of the more successful methods to mislabelling of data. For testing, we randomly chose



(a) Anger and Frustration

(b) Happiness and Excitement

(c) Happiness and Anger

(d) Sadness and Anger

(e) Excitement and Frustration

(f) Happiness and Sadness

Figure 5.10: Positions of images in the space of the first three principal components of some of the transformed emotion clusters of full face data in 4D space. All the figures are in 3D, but to get a clear view the viewpoint has been rotated by setting the azimuth and elevation to some suitable value. The figure needs coloured print to differentiate between different clusters.

2,000 frames of each of the six emotions to form a total set of 6,000 frames for each of the two actors separately.

The six classes that we used in our dataset were neutral, anger (ang), frustration (fru), happiness (hap), excitement (exc), and sadness (sad). However, we noticed that anger and frustration had a significant overlap, as did happiness and excitement. This has also been reported in the literature [25]. We therefore also used four classes: neutral, anger/frustration, happiness/excitement, and sadness, which is a significantly easier problem.

Fig. 5.11 shows the results on all images in the female test set for the six automatic methods of emotion recognition. While in the training set only images where all three experts agreed were used, the testing set contains most of the images where either two of the experts agreed or all three provided different assessments, often because one or more of them gave two labellings (such as anger/frustration). The label assigned for evaluation was based on the majority label out of the more than three labels attached to the frame by the three experts.

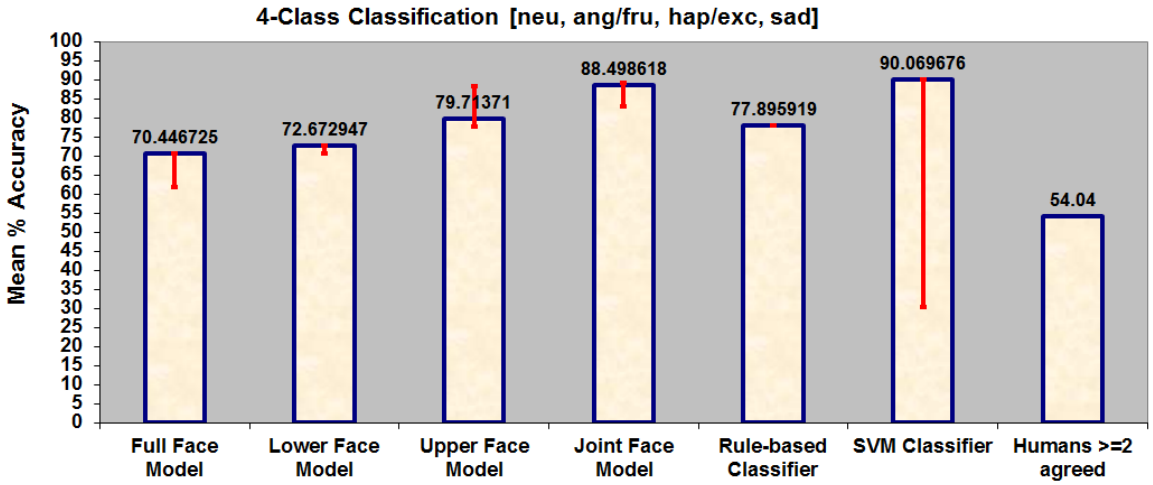
Fig. 5.11(a) shows the results of the four class classification problem. It can be seen that the *joint-model* returns almost 89% accuracy, which is consistently better than the *full* (70.45%), *upper* (79.71%), and *lower* (72.67%) face models, as well as *rule-based* PCA classifier (77.90%). On the female dataset (with no mislabelled frames), the SVM classifier seems to outperform other models, but it does not show consistent performance and remains unpredictable, especially for the data with high levels of mislabelling. In order to compute the difference in the *means* of joint model and SVM classifier, we perform the *t-Test: Paired Two Sample for Means* [231] on the results of these two models. The test suggests that there is no significant difference between these two results ($p > 0.05$).

Fig. 5.11(b) shows the results of classification to six emotion classes as well as for agreement between the three human observers for the female dataset. The classification into the six emotion classes is a difficult problem due to the significant overlap of anger and frustration as well as the happiness and excitement data clusters. Mostly the frames labelled as angry by human experts are classified as frustrated by the classifiers and vice-versa. The same is true for the frames labelled as happy and excited, which significantly decreases the performance accuracy. Fig. 5.12 shows the results on all images in the male test set for the six automatic methods of emotion recognition as well as for agreement between the three human observers for the male test set. For the four class classification problem, the *joint-model* outperforms other classifiers while for the six class classification, the performance of *joint-model* and that of the *rule-based* PCA classifier is not significantly different.

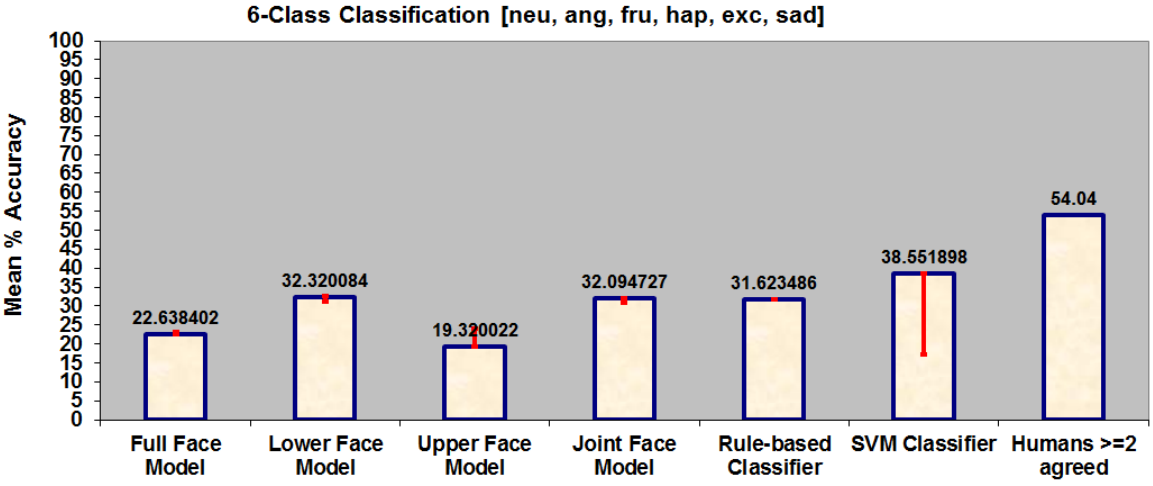
Table 5.4 lists the mean accuracy and standard deviation of the joint face model, full, upper, and lower face models, SVM Classifier, and the rule-based classifier on both 4 and 6 emotion classes for both male and female test sets. It shows that the joint face model significantly improves the accuracy, especially in the case of 4 class problem, as compared to other methods.

	Female 4-class mean(std)	Female 6-class mean(std)	Male 4-class mean(std)	Male 6-class mean(std)
Joint Face Model	88.50(3.44)	32.10(0.76)	70.18(2.77)	45.19(1.02)
Full Face Model	70.45(4.99)	22.64(0.28)	63.34(4.21)	41.74(0.51)
Upper Face Model	79.71(5.71)	19.32(2.80)	69.83(1.61)	41.12(1.05)
Lower Face Model	72.67(1.19)	32.32(0.82)	51.29(1.74)	33.88(1.47)
SVM Classifier	90.07(34.48)	38.55(12.26)	35.46(5.47)	35.39(5.68)
Rule-based Classifier	77.90(0)	31.62(0)	66.46(0)	47.76(0)

Table 5.4: Mean accuracy and standard deviation of the joint face model, full, upper, lower face models, SVM Classifier, and the rule-based Classifier on both 4 and 6 emotion classes.

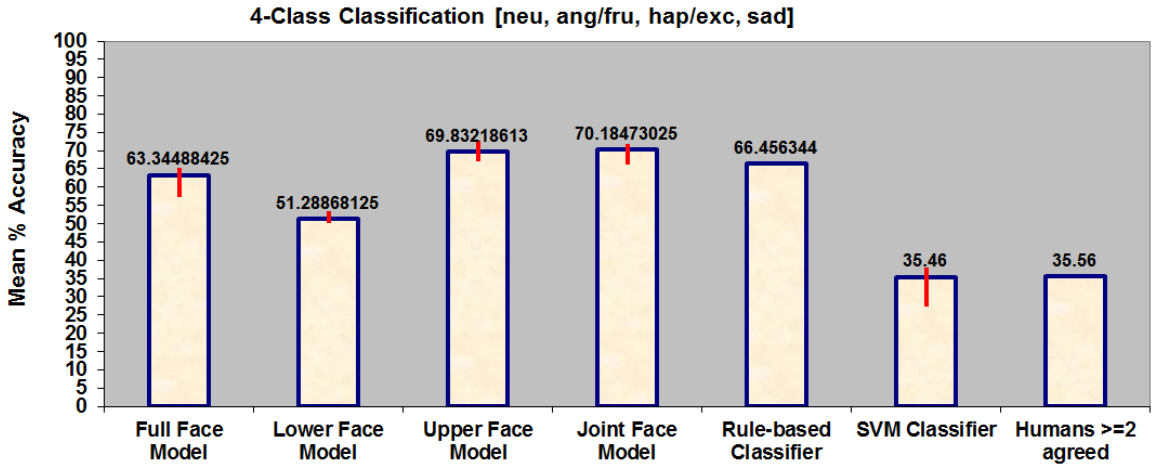


(a) 4-class classification (neu, ang/fru, hap/exc, sad)

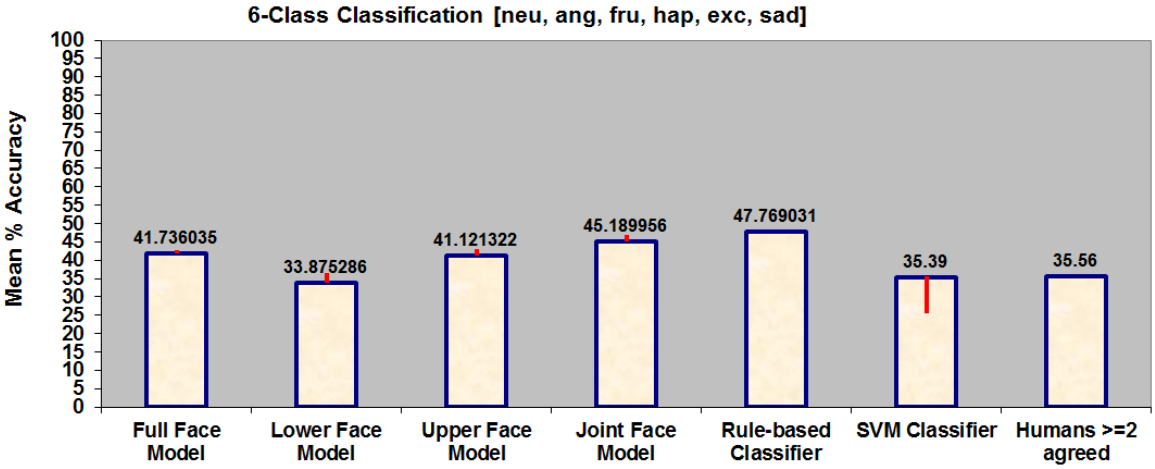


(b) 6-class classification (neu, ang, fru, hap, exc, sad)

Figure 5.11: The comparison of the mean accuracy of the four shape models, the rule-based classifier, and the SVM-based classifier on the female testset. Lines mark one standard deviation.

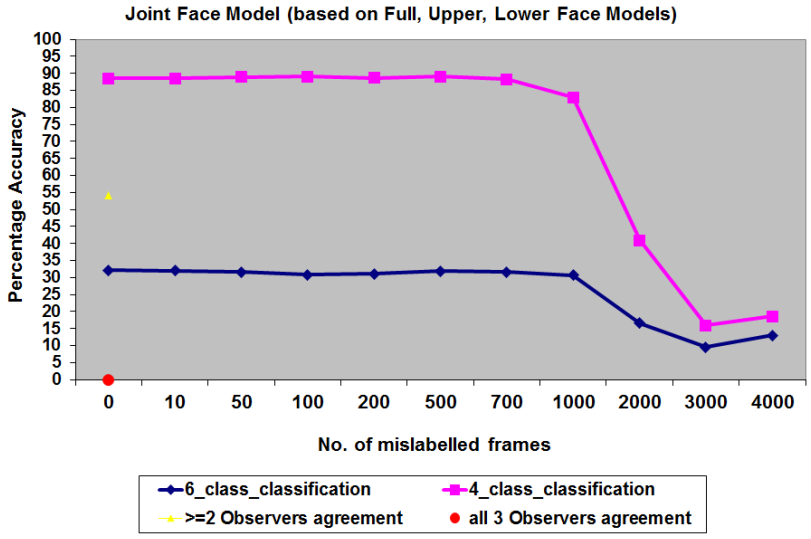


(a) 4-class classification (neu, ang/fru, hap/exc, sad)

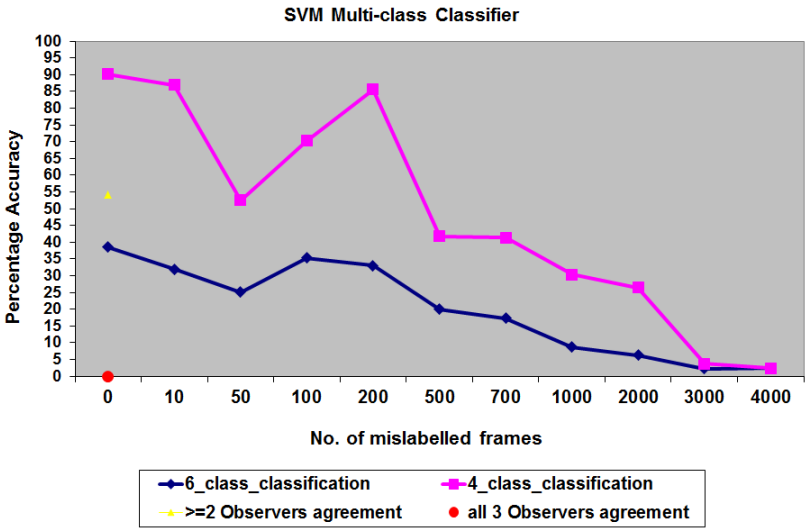


(b) 6-class classification (neu, ang, fru, hap, exc, sad)

Figure 5.12: The comparison of the mean accuracy of the four shape models, the rule-based classifier, and the SVM-based classifier on the male testset. Lines mark one standard deviation.



(a) The Joint Face Model



(b) SVM Classifier

Figure 5.13: The robustness of the joint face model and the SVM-based classifier to mislabelled training data.

The proposed joint model performs better than the individual human observers, who often assign more than one emotion label, despite the fact that the human observers had a whole lot of other information like voice, head, and hand motion along with the facial expressions, while the system has to classify on the basis of just the motion of facial markers.

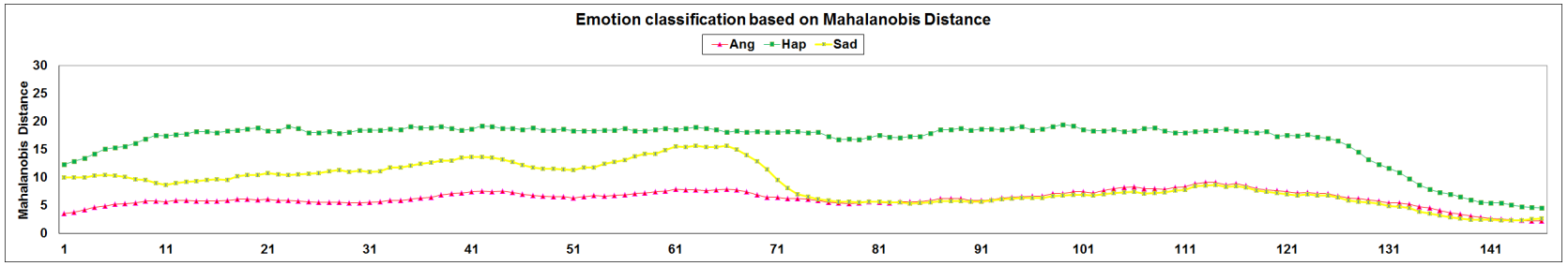
5.5.1 Robustness to mislabelled data

As mentioned in Chapter 4, there are two problems with the training data. First, the human experts were inconsistent and we were worried that the labels of the training data were compromised by errors. Second, the human experts labelled only utterances, and each utterance had many frames within it. The emotion label was attached to all frames in the utterance, but there is no guarantee that they all show one consistent emotion. In order to test the robustness of our system to data mislabelling, we manually mislabelled some proportion of the training data and ran the experiment again. The mislabelling was performed by choosing frames at random and then changing their label. This was done keeping in mind the possible errors that were likely to occur, so angry could become frustrated, frustrated could become angry or sad, neutral could become sad, and so on. We created datasets where 0.25%, 0.5%, 1.25%, 2.5%, 5%, 17.5%, 25%, 50% and 100% of the data were mislabelled.

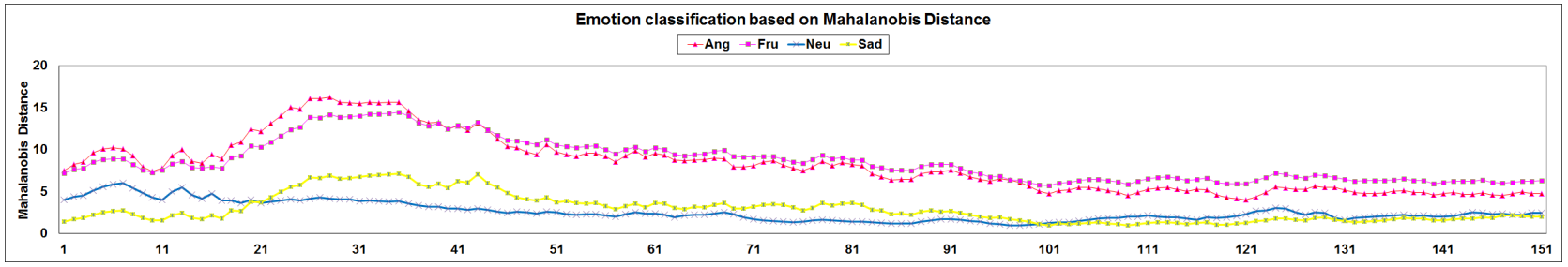
We tested and compared all above mentioned models including *full*, *upper*, and *lower* face models; proposed *joint-model*; *rule-based* PCA classifier and *SVM* classifier on gradually increasing mislabelled training data using both 6 and 4 emotion classes. Fig. 5.13 presents the performance of proposed joint model and SVM classifier on the female dataset. The graphs show that the system is resistant until 25% of the training data (i.e., 1000 out of 4000 frames) is mislabelled for the joint model, however the performance of SVM classifier is not consistent.

5.6 Discussion

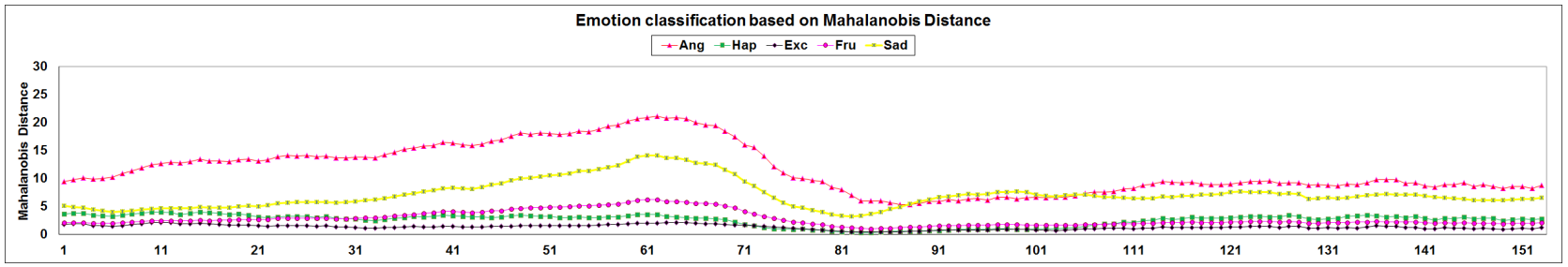
This chapter has presented our direct classification approach for discrete basic emotion recognition. Based on the observations that some emotions are better expressed by the upper part of face and some by the lower part, we proposed to develop separate



(a) The Mahalanobis Distances of testing frames (ground-truth: ang/fru, ang/fru, sad) to the happiness, anger, and sadness clusters



(b) The Mahalanobis Distances of testing frames (ground-truth: sad, sad/neu, neu) to the neutral, sadness, anger, and frustration clusters



(c) The Mahalanobis Distances of testing frames (ground-truth: exc, hap, hap/fru) to the anger, happiness, excitement, and frustration clusters

Figure 5.14: Some examples of emotion blends.

shape models for upper and lower regions of face. Our joint face model worked well, producing robust results which are more accurate than the individual human observers, who assigned more than one emotion label to each turn. We compared the results of our system with the majority label out of more than three labels assigned by three human evaluators.

As discussed in Chapters 2 and 3, giving a discrete name to emotions is not always possible, since in daily life we come across a lot of complex emotional states that cannot be given a discrete name. In other words, sometimes the difference between two or more emotions is so subtle that it becomes difficult to finely differentiate between them. The same thing we observed in our data. Consider Fig. 5.14(a) showing Mahalanobis distance of the testing frames in sequence with time to each of the clusters of anger, sadness, and happiness. The ground truth label of this utterance is anger/frustration, anger/frustration, sadness by three observers. It is clear that the distance of the last few testing frames to sadness and anger is so small that it seems inappropriate to assign a single label based on the minimum distance criteria. The unknown emotion seems to be a blend of sadness and anger/frustration. In Fig. 5.14(b), the distance of the testing frames (ground-truth: sad, sad/neu, neu) from sadness and neutral is swapping after small intervals of time. Similarly, in Fig. 5.14(c) the distance of the testing frames (ground-truth: exc, hap, hap/fru) to the happiness, excitement, and frustration clusters is almost equal. Due to this ambiguity, human evaluators assigned more than one categorical label to each turn.

Motivated by these thoughts, the next chapter address the problem of spontaneous emotion recognition by mapping them into the emotion space. Instead of classifying the emotions to one emotion category based on the best-matched criteria, we would be adding the effect of all basic emotions for the recognition of complex emotions. The effect of intensity will also be considered since in Fig. 5.14(b) it seems that the

subject is expressing sad emotion with low intensity (close to neutral state). Due to the robustness to mislabelled data and satisfactory recognition rate, the joint face model of basic emotion recognition is a reliable approach to be extended further to complex emotion recognition.

Chapter 6

Statistical Modelling of Complex Emotions using Mixtures of von Mises Distributions

6.1 Introduction

As discussed in the previous chapter, it is not always possible to assign a single categorical label to complex emotions, which are mixtures of two or more basic emotions. It was also seen in Chapter 2 that standard emotion spaces are circular, which means that we cannot use linear statistical models to describe emotion data. In this chapter, a statistical modelling technique for complex emotion recognition is provided by mapping them into the activation-evaluation space. The chapter starts with the discussion of circular data analysis that illustrates why the standard statistical techniques are inappropriate to analyse circular data. The efficacy of the proposed technique is demonstrated with experiments comparing the estimated results to those obtained by the psychological studies, as well as to the ground truth data.

6.2 Analysis of Circular Data

Angular observations may be regarded as observations on a circle of unit radius. A single observation θ measured in radians ($0 < \theta \leq 2\pi$), is then a unit vector, and the data can be described as circular data or directional data [140]. The Cartesian coordinates of the point at the end of the vector from the origin are $(\cos\theta, \sin\theta)$, while the polar coordinates are $(1, \theta)$.

Angular observations arise in many different contexts, e.g., studies of animal navigation in biology, seasonal fluctuation data in medicine, wind direction data in meteorology, and dip and declination data in geology [87]. From psychological studies (as discussed in Section 2.5), we know that emotions are related to each other in a circular manner [174], and the distance from an emotional state to the origin of the activation-evaluation space can be interpreted as the intensity of emotion [191]. The activation-evaluation space is a disk of potential emotions, and so a random variable describing emotional data is a circular random variable.

Due to the circular geometry of the sample space, the standard statistical techniques cannot be used to analyse emotion data. Suppose, $x_1, x_2, \cdots, x_n$ are n independent observations on a circle of unit radius, such that $0 < x_i \leq 2\pi$ where $i = 1, 2, \cdots, n$. The mean $\bar{x}$ of these observations cannot be calculated as $\frac{1}{n}\sum_{i=1}^{n} x_i$. The reason for this is that a circle does not have a fixed starting or ending point because the beginning and end of the circle coincides: $0° = 360°$. For example, consider a sample of size 2 consisting of two angles 1° and 359°. The arithmetic mean of these two angles is 180°, which is different from the geometrical mean of 0° [140].

It is therefore appropriate to regard angular observations on the circle as unit vectors in the plane, and take the geometric mean of these vectors, referred to as the resultant vector. Fig. 6.1 (taken from [19]) illustrates the resultant vector direction and length obtained by adding three different angles on the circle. Fig. 6.1(a)

illustrates the mean of three angles $(60°, 180°, 300°)$ yielding a resultant vector length of zero (we cannot see the resultant vector of length zero, it resides at the origin) because the points are exactly uniformly spaced around the circle. Fig. 6.1(b) represents the mean of three angles $(120°, 180°, 240°)$ yielding a resultant vector length of $\frac{2}{3}$, with a mean direction of 180°. Fig. 6.1(c) represents the mean of three angles $(150°, 180°, 210°)$ yielding a resultant vector length of 0.9107, with a mean direction of 180°. All the angles are taken from the horizontal zero-direction. It is clear from these figures that only if the observed directions on the circle are tightly clustered around the mean direction will the length of mean resultant vector be almost 1.

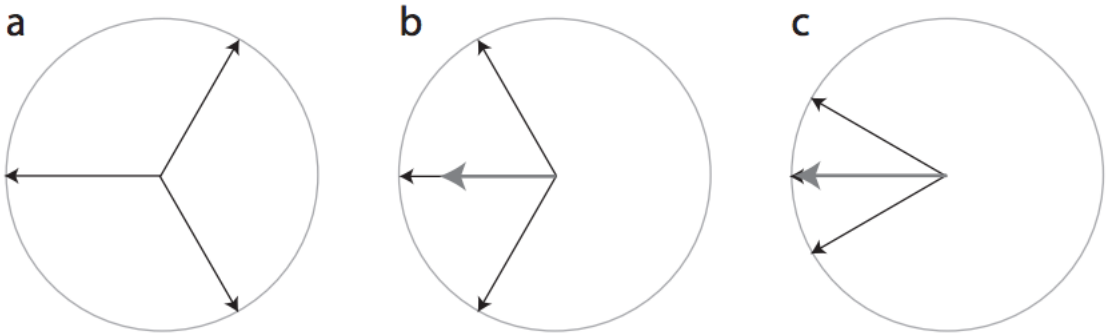


Figure 6.1: Illustration of the resultant vector direction and length (in grey) obtained by vector addition of (a) $60°, 180°, 300°$, (b) $120°, 180°, 240°$, and (c) $150°, 180°, 210°$. All the angles start from the horizontal zero-direction. The light grey circle is the unit circle. In (a), we cannot see the resultant vector of length zero, it resides at the centre of the circle. This picture is taken from [19].

6.2.1 Circular distribution

A circular distribution is a probability distribution whose total probability is concentrated on the circumference of a unit circle [112]. A given function f is the probability density function of an absolutely continuous circular distribution if and only if:

$$f(\theta) \geq 0, \quad -\infty \leq \theta \leq \infty \tag{6.1}$$

$$f(\theta + 2\pi) = f(\theta), \quad -\infty \leq \theta \leq \infty \tag{6.2}$$

and,

$$\int_0^{2\pi} f(\theta) d\theta = 1. \tag{6.3}$$

In this case, θ will be described as a continuous random variable [140].

6.2.2 Statistical approaches to modelling circular data

The majority of the circular probability distribution models are derived from the standard linear models or are their circular analogues. The main distributions among the continuous circular distributions are the uniform distribution, and the von Mises distribution. Any linear distribution can be wrapped around the circumference of the circle of unit radius. Such distributions are called wrapped distributions. Suppose x is a random variable on the line, the random variable x_w of the wrapped distribution is given by $x_w = x(\mod 2\pi)$. The most common among the wrapped distributions are the wrapped normal distribution and the wrapped Cauchy distribution [82, 112, 140].

The von Mises probability distribution function (pdf) is the most common model for symmetric uni-modal samples of circular data and is a circular analogue of the standard normal distribution [140]. This distribution was introduced by Richard von Mises in 1918, in order to study the deviation of measured atomic weights from integral values. In directional statistics its importance is almost the same as that of the normal distribution on the line [82, 112, 140]. The von Mises distribution is a popular choice due to the possibility of calculating the maximum likelihood parameter estimations and the fact that it is easy to interpret.

The von Mises distribution depends on two parameters: the mean direction μ and the concentration parameter K, which are analogous to μ and σ^2 of the standard normal distribution. For $K = 0$, the von Mises distribution reduces to the uniform

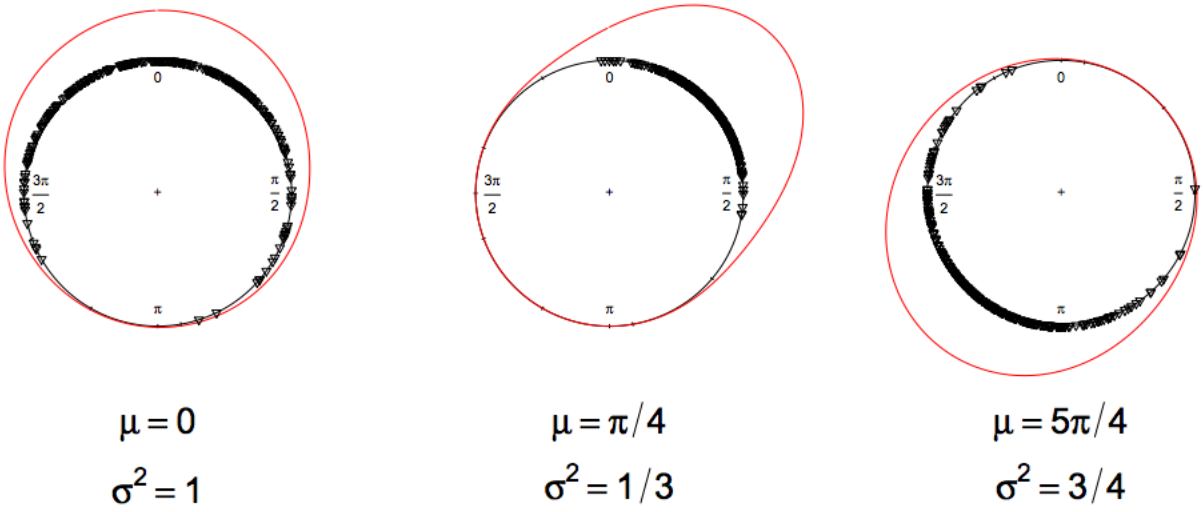


Figure 6.2: The probability density plots for three different sets of datapoints modelled using wrapped normal distributions with selected parameters μ and σ^2. The red circles show the density estimate based on the observed data. Picture taken from [87].

distribution, while for large K, the distribution becomes the wrapped normal distribution [36, 64]. For $K \to \infty$, the distribution becomes concentrated at a single point $\theta = \mu$. The von Mises distribution will be described in detail in Section 6.3.2.

The wrapped normal distribution has the same parameters μ and σ^2, as the standard normal distribution. The wrapped normal distribution is also uni-modal and symmetric about $\theta = 0$. Fig. 6.2 shows the probability density plots for three different sets of datapoints modelled using wrapped normal distributions with selected parameters μ and σ^2. The red circles show the density estimate based on the observed data. The distribution is highly sensitive to the parameter of σ^2.

In the case of data with heavier tails, the wrapped Cauchy distribution may be considered. It is obtained by wrapping the linear Cauchy distribution around the circumference of a unit circle. Its mean and variance are undefined. Fig. 6.3 shows a probability density plot for 100 datapoints randomly drawn from a wrapped Cauchy distribution centered at π on the unit circle. The 'heavy tails' are expressed as a large number of points that fall far away from π. The probability density function of the

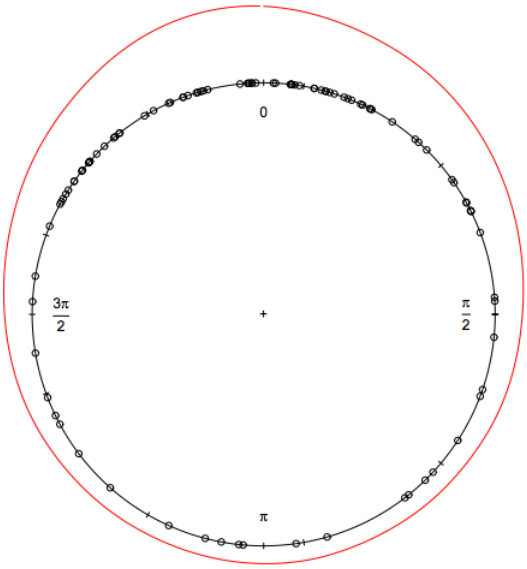


Figure 6.3: The probability density plot for 100 datapoints randomly drawn from a wrapped Cauchy distribution. The red circle shows the density estimate based on the observed data. Picture taken from [87].

wrapped Cauchy distribution $c(\theta, R)$, is a function of θ and the resultant vector R is the Poisson kernal [140].

6.2.3 Selection of an appropriate circular distribution

Based on observations of the datapoints mapped into activation-evaluation space (using the method described in the next section), we believe that our data is approximately normal in a local neighbourhood. Therefore, the uniform probability distribution is not suitable to describe this data. Among the normal distributions, the von Mises distribution and the wrapped normal distribution closely approximate each other. As already mentioned, for large K, the von Mises distribution is transformed to the wrapped normal distribution with mean μ and variance $\frac{1}{K}$. Also, there is satisfactory agreement for moderate values of their respective concentration parameters [205].

Like the linear normal distribution, for the von Mises distribution the maximum likelihood estimation of parameters, which is relatively complicated for other circular distributions, is simple. Also, it maximises the entropy for a fixed circular mean and variance. On the other hand, the wrapped normal distribution is a natural adaptation of the linear normal distribution to the circle. Most of the properties of the wrapped normal distribution hold true for the von Mises distribution and vice-versa, so we choose the one with simpler maximum likelihood estimation of parameters.

We have chosen to use the von Mises probability distributions to represent the distributions of six basic emotions, since it allows a feasible maximum likelihood estimate of its parameters. Also, we will use the von Mises mixture model to interpolate complex emotions. The von Mises mixture model will be described in Section 6.3.2. We have used the *CircStat* toolbox [19] of Matlab for the statistical analysis of circular data.

6.3 Description of Our Method

We introduce a flexible mixture model to recognise complex emotions on the basis of known basic emotions. Following the assumption that complex emotions can be conceived of as mixtures of basic emotions [177], we hypothesise that it may be possible to develop a mixture model that combines each basic emotion in an appropriate amount to recognise and represent complex emotions. The proposed model is based on the activation-evaluation space, which is the most widely-used representation of emotions in psychological studies [178]. The activation-evaluation space has been discussed in Section 2.5.2.

6.3.1 Mapping emotions to the activation-evaluation space

We have calculated the Mahalanobis distance of each test frame to each of the six basic emotions using the shape models (using the method described in Section 5.2). There are two steps required to map the representation of the facial points of an image frame into the activation-evaluation space: represent the basic emotions as points within that space, and then position each frame (using the six distances to the basic emotions). The first of these steps uses the training data, which is assumed to represent each of the six basic emotions (all three experts agreed on their labels), while the second uses the testing data.

As well as giving the emotion class label, the human experts annotated each utterance with valence and activation values (as a self-assessment manikin (SAM) score between 1 and 5; we rescaled these values to [-1,+1]). For the 6,000 frames of each emotion this represented 10-15 utterances, so we had 10-15 values for three expert annotations of each emotion. We transformed these into polar coordinates (which corresponds to intensity of emotion in the radial direction and particular emotion in the angular direction) and then computed the mean average of the 30-45 values for each emotion.

For the radial component the linear statistical mean average is correct, but in order to average angles, it is not appropriate (as described in Section 6.2). Therefore, we chose to calculate the geometric mean to get the angular position of each of the basic emotions in the activation-evaluation space. The computation of the mean direction and magnitude of the resultant vector for each of the six emotions separately is as follows:

$$\bar{V} = \frac{1}{n}\sum_{i=1}^{n} val_i, \quad \bar{A} = \frac{1}{n}\sum_{i=1}^{n} act_i \tag{6.4}$$

where n is the number of utterances of each emotion.

$$\mu = \begin{cases} \tan^{-1}(\bar{A}/\bar{V}) & \bar{A} > 0, \bar{V} > 0 \\ \tan^{-1}(\bar{A}/\bar{V}) + \pi & \bar{V} < 0 \\ \tan^{-1}(\bar{A}/\bar{V}) + 2\pi & \bar{A} < 0, \bar{V} > 0 \end{cases} \tag{6.5}$$

$$\bar{R}^2 = \bar{V}^2 + \bar{A}^2 \tag{6.6}$$

μ is the mean direction and $\bar{R}$ corresponds to the intensity (in terms of valence and activation) in the activation-evaluation space.

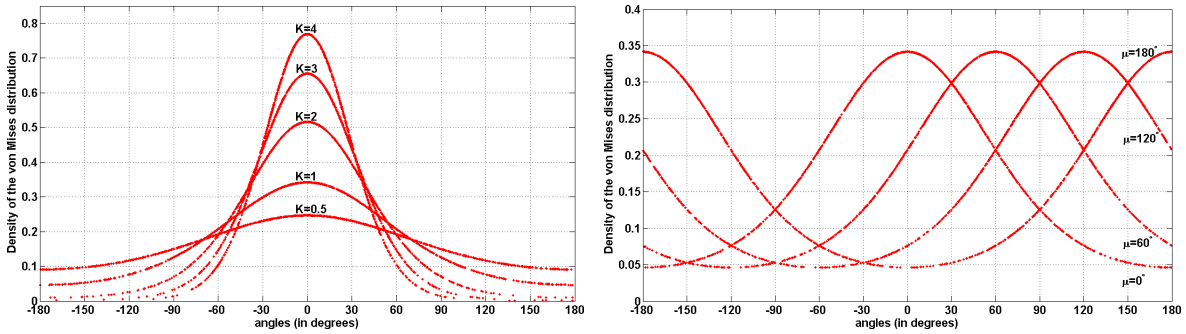
This gave us locations for the six basic emotions (including neutral). Each test frame is assumed to be a combination of the basic emotions, and so we needed to calculate the weighted average of basic emotions, where the weights correspond to the classification confidence of test frames for each basic emotion. We modelled the distribution of each basic emotion as a von Mises distribution and constructed a mixture model of them to calculate the weighted average of basic emotions for each test frame. This is described in Section 6.3.2. In the work described in this chapter the emotions corresponding to each utterance are mapped into the activation-evaluation space frame by frame; however, in the next chapter we will describe the computational analysis of continuous emotion trajectories to understand emotion dynamics.

6.3.2 von Mises Mixture Model

A circular variable θ is said to have a von Mises distribution if the probability density function is given by:

$$m(\theta; K, \mu) = \frac{1}{2\pi I_0(K)} e^{[K\cos(\theta-\mu)]} \tag{6.7}$$

where $0 \leq \theta < 2\pi$, $K > 0$ and $0 \leq \mu < 2\pi$.



(a) $\mu = 0°$ and $K = 0.5, 1, 2, 3, 4$ (b) $K = 1$ and $\mu = 0°, 60°, 120°, 180°$

Figure 6.4: The effect on the density of von Mises distribution for (a) fixed mean direction μ and varying concentration parameter K and (b) fixed concentration parameter K and varying mean direction μ.

The parameter μ is the mean direction and K is the concentration parameter, which is analogous to the (inverse) variance: the density at the mode depends on e^{2K} and the larger the value of K, the greater is the clustering around the mode. Figure 6.4(a) shows the effect on density of the von Mises distribution for a fixed mean $\mu = 0°$, and varying $K = \frac{1}{2}, 1, 2, 3, 4$, while figure 6.4(b) shows the effect on the density for fixed $K = 1$, and varying $\mu = 0°, 60°, 120°, 180°$. The distribution is uni-modal and symmetric about μ. $I_0(K)$ is a normaliser to turn this into a probability density function and consists of a modified Bessel function of the first kind of order zero [27]:

$$I_0(K) = \sum_{r=0}^{\infty} \frac{1}{r!^2} \left(\frac{1}{2} (K)^{2r} \right) \tag{6.8}$$

Although each emotion class is uni-modal, we cannot fit one von Mises distribution to the full data as it is the mixture of six different emotion classes. Such multi-modal distributions may be regarded as mixtures of uni-modal distributions [140]. We used

a finite mixture model of six uni-modal von Mises distributions, given by:

$$M = \sum_{j=1}^{6} \omega_j m_j(\theta) \tag{6.9}$$

where ω_j are non-negative weights that sum to one. We have already calculated the mean direction (μ_j) of each of the six reference emotions in the space using Eq. (6.5), and the methods of estimating K_j and ω_j are described in the following section.

6.3.3 Estimating the Parameters of the Mixture Model

There are several ways to estimate the parameters on which the mixture model depends [140]. We have used the usual maximum likelihood estimate for K_j. However, the weights ω_j of each emotion model are estimated by using the distances to the six emotions calculated by the shape models.

6.3.3.1 Estimating the concentration parameter

The concentration parameter K_j is estimated by using the Fisher equation [82]:

$$\hat{K}_{ML} = \begin{cases} 2\bar{R} + \bar{R}^3 + 5\bar{R}^5/6 & \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + 0.43/(1-\bar{R}) & 0.53 \leq \bar{R} < 0.85 \\ 1/(\bar{R}^3 - 4\bar{R}^2 + 3\bar{R}) & \bar{R} \geq 0.85 \end{cases} \tag{6.10}$$

$\hat{K}_{ML}$ may be biased if the sample size (n) and $\bar{R}$ are small (specially when $\bar{R} < 0.45$). For this reason, if $n \leq 15$, the following estimate is to be preferred:

$$\hat{K} = \begin{cases} \max(\hat{K}_{ML} - 2(n\hat{K}_{ML})^{-1}, 0) & \hat{K}_{ML} < 2 \\ (n-1)^3 \hat{K}_{ML}/(n^3 + n) & \hat{K}_{ML} \geq 2 \end{cases} \tag{6.11}$$

This is a standard approach for estimating the concentration parameter [82, 140].

6.3.3.2 Estimating the weights

We have already calculated the Mahalanobis distance of each test frame to each of the six basic emotions using the shape models and retained the minimum distance of the three models for each emotion. We want to position each test frame in activation-evaluation space using the positions of the basic emotions. However, the Mahalanobis distance is an unsigned quantity and so we do not know the direction between the test frame and the mean of each of the clusters of basic emotions. Since we have assumed that each emotion lies along a radial line in the activation-evaluation space we want to compute the intensity of each of the basic emotions as a component of the complex emotion. We did this starting at the position of the basic emotion and then by applying a simple rule to move along that radial line: if the distance of the test frame from neutral is less than the mean of a particular emotion, then the distance of the test frame from that emotion should be towards neutral i.e., its intensity decreases and comes close to neutral and vice-versa. We convert these distances to weights (ω_j) by reciprocating their values.

6.4 Experimental Results

Fig. 6.5 plots the locations of the basic emotions in the activation-evaluation space using the estimated mean positions of the training set. We observed that the estimated directions are quite close to those specified by Whissell in [221], except that of neutral state. This seems to be because mostly the human evaluators misinterpreted neutral as sadness and assigned a low activation value to the neutral state. According to the values listed by Whissell, neutral lies close to the centre of circle that

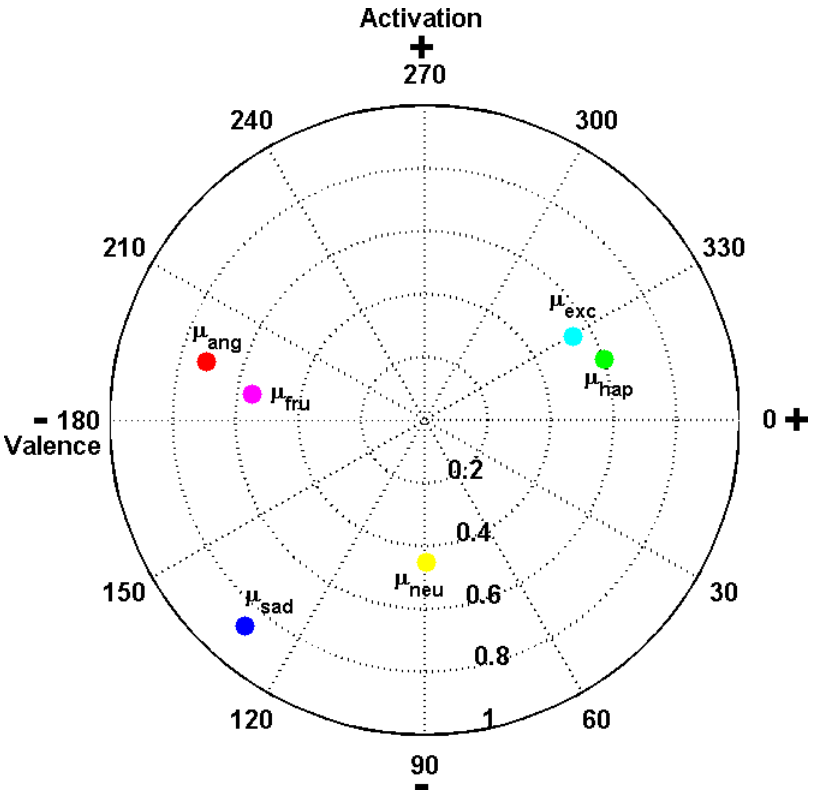


Figure 6.5: The position of each basic emotion (based on the training set) in the activation-evaluation space.

corresponds to both valence and activation close to zero (see Table 6.1 for numerical values). To correct the position of the neutral state, we set the intensities of points labelled as neutral to be close to zero (near the centre) in the activation-evaluation space. This shift does not effect the position of other emotions because the neutral state represents the 'no-emotion' state whose position should be uncorrelated with that of all emotional states.

Based on these positions for the basic emotions we were now able to compute the parameters of the mixture model and test it using initially single utterances with only one labelled emotion, and then full conversations with several emotion transitions. Fig. 6.6(a) shows the von Mises probability distributions of all emotions in the mixture model. The x-axis shows the emotion directions (angles in degrees) (see Table 6.1 for

Table 6.1: Angular values from Whissell's study and those estimated by the models.

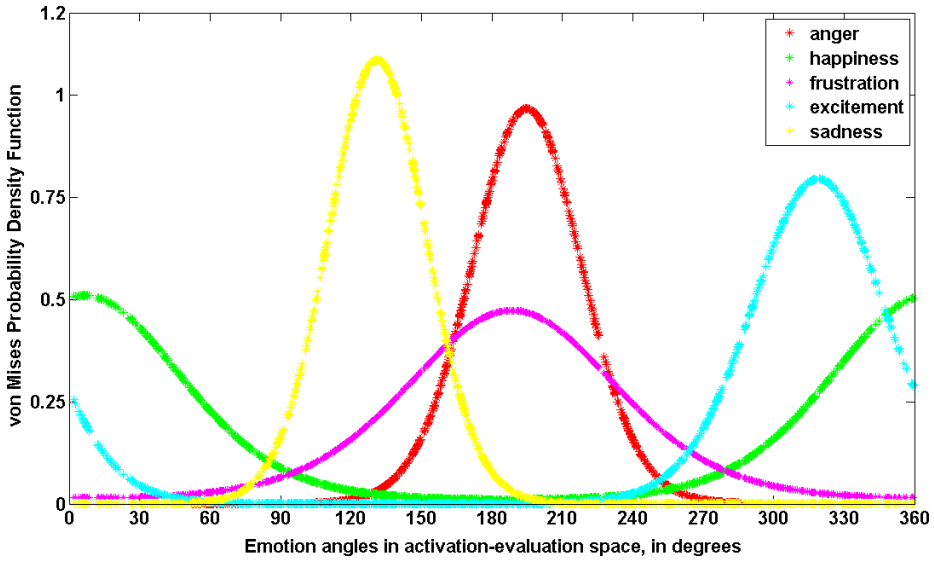
Emotions	Whissell's Angles (in degrees)	Estimated Angles (in degrees)
Sadness	108.5	131.10
Frustration	200.6	188.48
Anger	212	194.84
Excitement	311	330.58
Happiness	323.7	341.45
Neutral	0	89.23 (inclined towards *very passive* state)

numerical values) and the y-axis shows the von Mises probability densities for each distribution in the mixture model. It is clear that there is a big overlap between the distributions of anger and frustration as well as of happiness and excitement. Sadness lies quite close to the anger/frustration distribution. This is to be expected based on the Whissell angles. Fig. 6.6(b) shows the von Mises probability distributions characterising one frame of an utterance labelled as [angry, angry, frustrated] by three human experts.

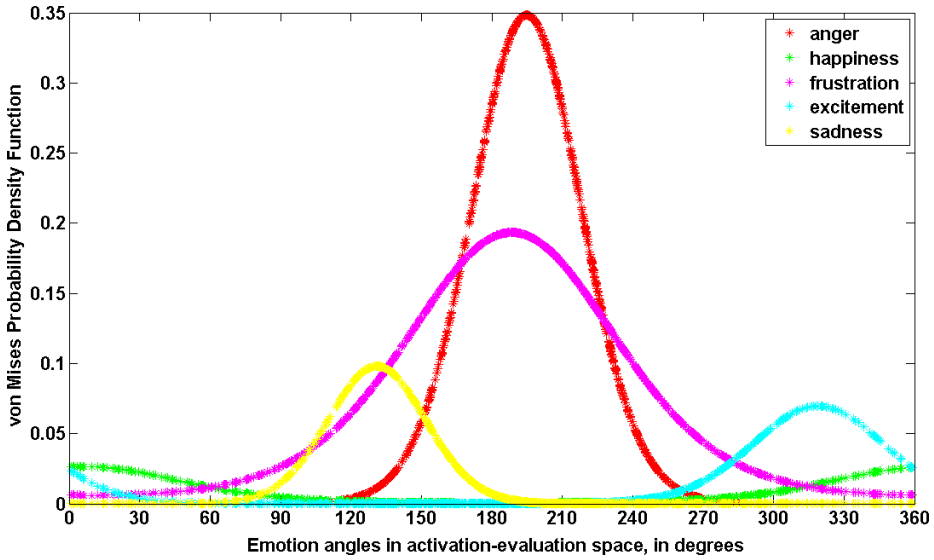
We use the valence and activation values assigned to each utterance by three human experts to estimate the ground truth direction and intensity of emotion associated with that utterance, which we can compare to our results. In order to measure the association between two circular variables (the ground truth direction and that estimated by the model), we have used a measure of circular sample correlation coefficient ($\rho_{c,n}$) [112]. If $(\alpha_1, \beta_1), \cdots, (\alpha_n, \beta_n)$ is a random sample, $\rho_{c,n}$ is defined as:

$$\rho_{c,n} = \frac{\sum_{i=1}^{n} \sin(\alpha_i - \mu) \sin(\beta_i - \nu)}{\sqrt{\sum_{i=1}^{n} \sin^2(\alpha_i - \mu)) \sin^2(\beta_i - \nu)}} \tag{6.12}$$

where μ and ν are the sample mean directions.



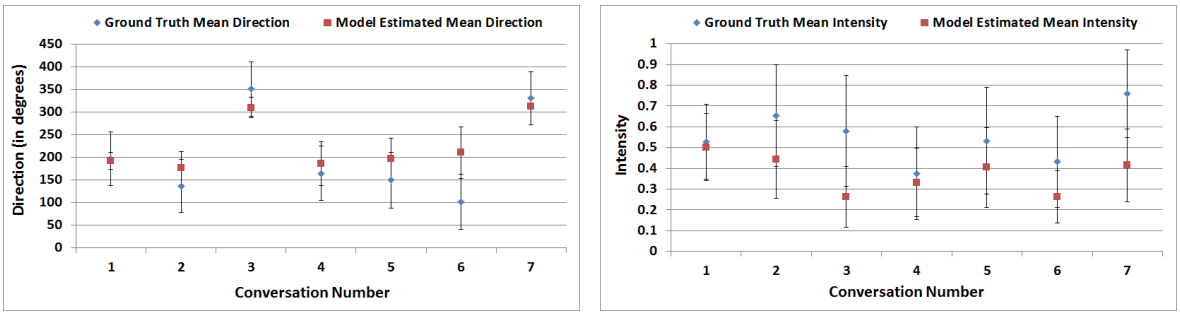
(a)



(b)

Figure 6.6: (a) von Mises Probability Distributions in the mixture model with unit weight, (b) von Mises Probability Distributions characterising one frame of an utterance labelled as [angry, angry, frustrated] by three human experts.

Only one ground truth measure of direction and intensity is available for each full utterance, while the model estimates the directions and intensity for each frame. In order to measure $\rho_{c,n}$, we generate a sample of (n=1000) random variables based



(a) (b)

Figure 6.7: The test of fit for (a) the mean ground truth *directions* and those estimated by the mixture model (b) the mean ground truth *intensities* and those estimated by the mixture model, for each of the seven conversations in the test set. Lines mark one standard deviation.

on the von Mises probability density functions for both distributions (ground truth and model estimation). $\rho_{c,n}$ shows a significantly high correlation between samples of ground truth direction and those estimated by the model for each conversation. We also applied the pairwise t-test on the intensity values and found that the intensity calculated by ground truth values and those estimated by the mixture model are not statistically different ($p > 0.05$). Fig. 6.7 presents the visual fit of ground truth mean directions and mean intensities with those estimated by the model for the test set consisting of seven different conversations.

In order to test the goodness of fit of two sample distributions (ground truth and model estimation), we also used Kuiper's test, which is a circular analogue of the Kolmogorov-Smirnov test [140, 218]. Let $F(\theta)$ denote the continuous cumulative distribution function (cdf) of each of the emotions in the mixture model separately and $S_n(\theta)$ be the ground truth cdf (referred to as the empirical cdf). The Kuiper's statistic is defined as,

$$V_n = D_n^+ + D_n^- \tag{6.13}$$

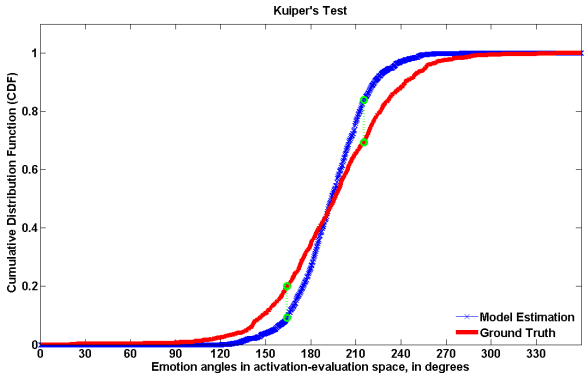
where

$$D_n^+ = \max[S_n(\theta) - F(\theta)], \quad D_n^- = \max[F(\theta) - S_n(\theta)] \tag{6.14}$$

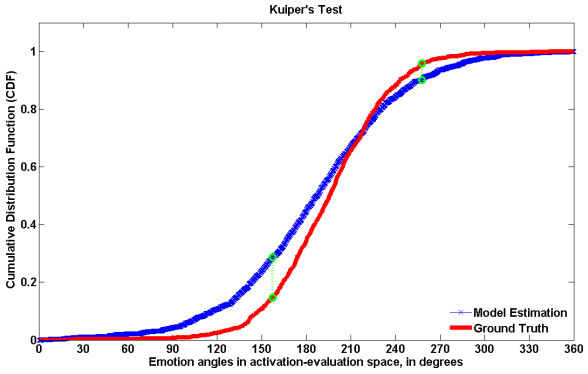
D_n^+ and D_n^- are the discrepancy statistics, where D_n^+ is the maximum vertical distance of $S_n(\theta)$ from $F(\theta)$ when the distance is measured above $F(\theta)$, while D_n^- is the maximum distance measured below $F(\theta)$. Both statistics D_n^+ and D_n^- depend on the choice of zero direction, but their sum (V_n) is invariant under rotation. This makes the Kuiper's statistic equally sensitive at the median as well as at the tails [125, 140].

Fig. 6.8 shows the model estimated cdf of each of the emotions in the mixture model separately and the empirical cdf for the first conversation in the testing set. It can be seen that the model estimated cdf for anger (Fig. 6.8(a)) as well as that of frustration (Fig. 6.8(b)) fits the empirical cdf well. However, the cdfs for happiness (Fig. 6.8(c)), excitement (Fig. 6.8(d)), and sadness (Fig. 6.8(e)) show quite high deviation from the empirical cdf. This suggests that the empirical distribution and the model estimated distributions of anger as well as frustration are not statistically different. On the basis of this, we may say that the first conversation in the testing set is an angry/frustrated conversation.

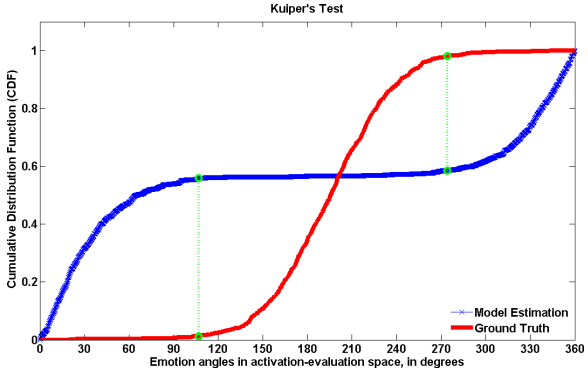
Fig. 6.9 shows the mapping of continuous emotions through time corresponding to a single utterance (Ses01F_script02_1_F007) in the activation-evaluation space. The colour variation represents time, ranging from red (dark in grayscale) to yellow (light in grayscale). The utterance is labelled as angry/angry/frustrated by the three human observers, which matches the observation well. The figure also plots the ground truth values of the mean direction and intensity and those estimated by the model. The analysis of continuous emotions through time in the activation-evaluation space is described in Chapter 7.



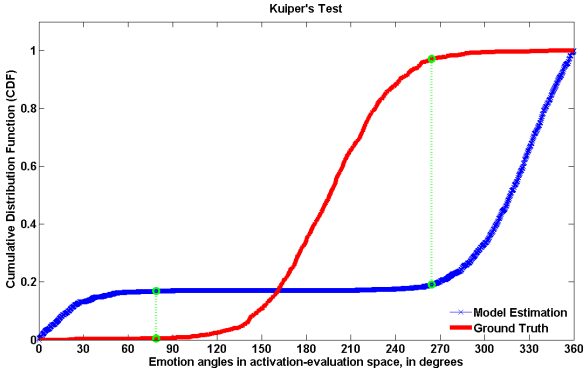
(a) cdf of anger



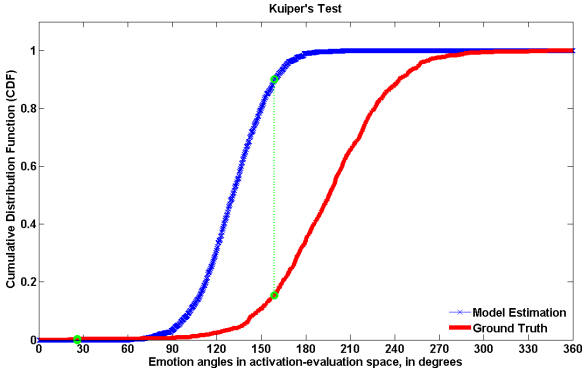
(b) cdf of frustration



(c) cdf of happiness



(d) cdf of excitement



(e) cdf of sadness

Figure 6.8: Illustration of the Kuiper's statistic. Blue line is model estimated CDF of (a) anger, (b) frustration, (c) happiness, (d) excitement, and (e) sadness, red line is an empirical CDF of the first conversation in the testing set, and the green line is the Kuiper's statistic.

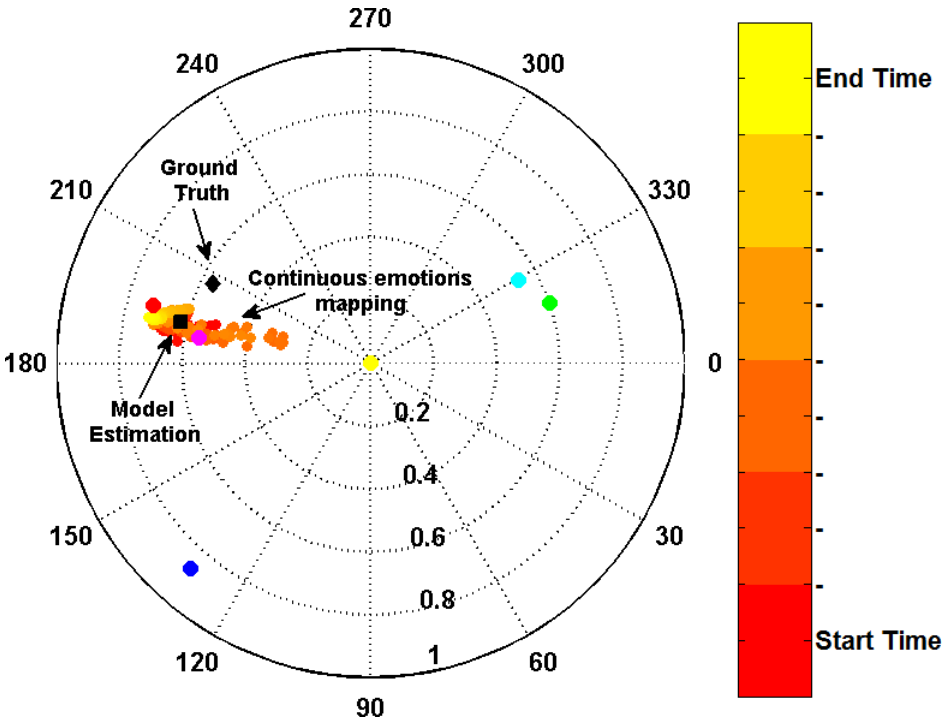


Figure 6.9: The mapping of continuous emotions during an utterance on the activation-evaluation space, along with the mean ground truth direction and intensity and those estimated by the model. The movement of emotions through time is represented by changing colour spectrum from dark/red (start) to light/yellow (end).

Figs. 6.7, 6.8, and 6.9 show that the proposed mixture model fits the data well, despite the underlying problems with the ground truth labelling (that is, the fact that there is only one label associated with each utterance, which lasts for many frames while the model estimates the values for each frame). Furthermore, all 'silent' frames are unlabelled in the conversations, while the model estimates the values for those frames as well. The intensity values do not fit as well as the directions because the small number of samples leads to high concentration around the mean as compared to the large number of frames in the testing set.

6.5 Summary

This chapter has presented a brief introduction to the circular data analysis along with a discussion of why standard algebraic operations and linear probability distributions are unsuitable for dealing with circular data. It also described some of the most widely-used circular probability distribution models and mentioned the similarities and differences between these models. Based on this discussion, we found that the von Mises distribution and the wrapped normal distribution are close approximation to each other. However, the von Mises distribution offers simpler maximum likelihood estimation of parameters, and is easier to interpret. Therefore, we chose to use the von Mises probability distribution to model six basic emotions, and the von Mises mixture model to describe complex emotions.

Using the von Mises mixture model, we presented a technique for the recognition and representation of complex emotions in the activation-evaluation space. The proposed method is based on the psychological assumption that complex emotions are comprised of mixtures of basic emotions. There is still debate among psychologists on the number of basic emotions and which emotions should be considered as basic, and of the six emotions that we have considered two (frustration and excitement) are candidate basic emotions. However, the proposed mixture model is quite flexible and can be applied to any set of basic emotions. We estimated the degree of similarity of each test frame to each of the basic emotions and project them into the activation-evaluation space using the von Mises mixture model.

In this chapter each continuous conversation is mapped into the activation-evaluation space frame by frame, but in the next chapter we will describe the computational analysis of continuous emotion trajectories to understand emotion dynamics. Emotions vary in intensity, flow, persistence with time, and their relationships with other emotions. By analysing the emotion dynamics through time, we try to seek the answers

about the 'common' paths between emotions, the smoothness of emotion trajectories, and how we travel along emotion flows. The computational analysis of emotion dynamics may be helpful for better understanding of emotion trajectories as well as in the development of more flexible models for emotion recognition, representation, and synthesis.

Chapter 7

Computational Analysis of Emotion Dynamics

7.1 Introduction

In this chapter we focus on the change in emotions over time. Research in psychology [185] and neuroscience [47] has shown that emotions change (more specifically, *fade*) over time due to either external stimuli or their natural progression. We focus on the transitions among six basic emotional states observed in the activation-evaluation space.

As in the previous chapters, we use facial expressions as the representative of the underlying emotional states. Just like the underlying emotions, facial expressions of emotions change over time, due to external stimuli and the way the face works. The facial dermal tissues comprise collagen (72%) and elastin (4%) fibres, which help resist deformation of tissues. Therefore, the facial tissues effected by the active facial muscles need to relax before stretching to another form (i.e., expressing another emotion) [207]. The movement of facial points among different emotional states is

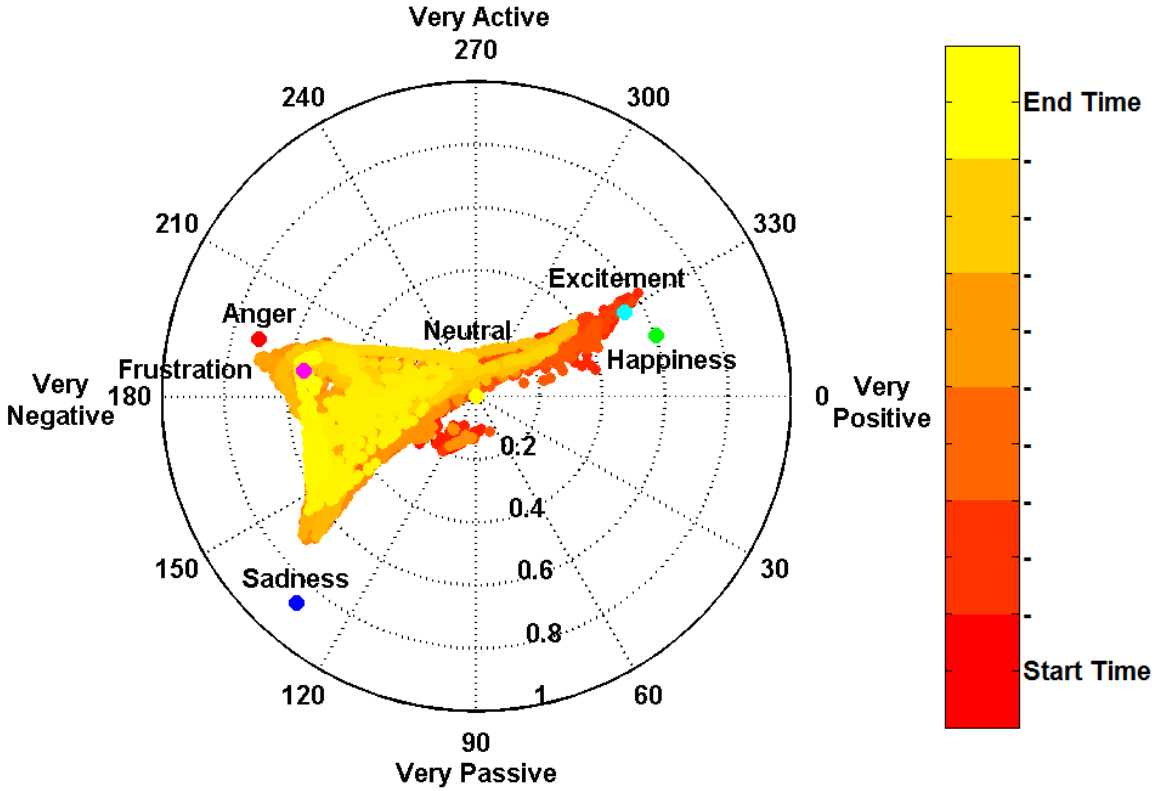


Figure 7.1: The mapping of continuous emotions during a conversation (Ses01F_impro01) into the activation-evaluation space. The movement of emotions through time is represented by changing colour spectrum from dark/red (start) to light/yellow (end).

analysed by tracking the paths of change in facial expressions through time. We do not consider the effect of external stimuli on the facial expressions of emotions.

7.2 Hypotheses

The paths of emotional expressions are observed in the activation-evaluation space. Fig. 7.1 shows the mapping of continuous emotions through time corresponding to an unsegmented conversation in the activation-evaluation space (using the methods described in Chapter 6). The colour variation represents time, ranging from red (dark in grayscale) to yellow (light in grayscale). The figure shows some sequence

information in the transition from one emotion state to another. Also, we already knew the mechanical properties of facial dermal tissues and that the emotions are related to each other in a systematic manner, as discussed in Section 2.5. All these facts and findings motivated us to computationally analyse the paths of emotions while transitioning from one state to another. We propose the following hypotheses about the emotion paths representation in the activation-evaluation space:

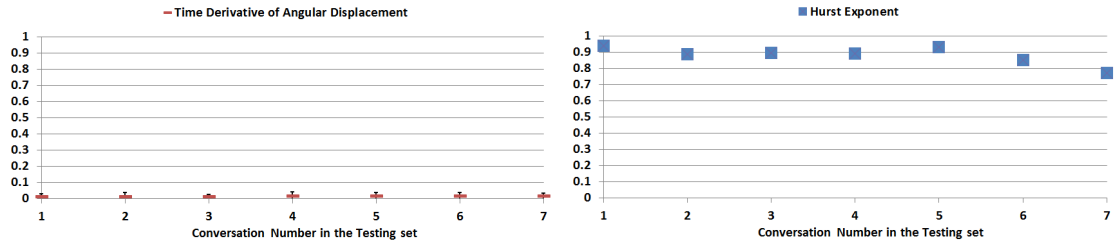
1. The paths among emotions form 'smooth' trajectories in the space.
2. If the end-point emotions are not positively correlated, then the path goes through the neutral state.
3. If the end-point emotions are positively correlated, then the path does not go through the neutral state.

Positively correlated emotions are those that lie close to each other (less than 90° divergence) in the activation-evaluation space. Uncorrelated emotions correspond to the emotions that lie at 90° divergence, while the negatively correlated emotions corresponds to the emotions which are 180° apart in the activation-evaluation space. These terms have already been described in Section 2.5.

7.3 Evaluation

Within the following section the processes used to test the proposed hypotheses are presented, together with the results of those tests.

Hypothesis 1: The paths among emotions form 'smooth' trajectories in the space. In order to test the first hypothesis, we first need to define 'smoothness' of an emotion trajectory. After extensive review of the literature, we could not find any



(a) (b)

Figure 7.2: Test of smoothness of emotion paths corresponding to each conversation in the testing set. For explanation, see the text.

standard definition. However, if we consider an emotion trajectory as a time series (sequence of values at successive time points following a non-random order), then its smoothness may be defined as a measure of its persistence with time. A random time series, e.g., Brownian motion, is not smooth, as it is not persistent with time. On the basis of this definition, we may say that if the points in the emotion space move in a predictable manner then the resulting trajectory is smooth/persistent with time.

We measure the smoothness of emotion points trajectories in the activation-evaluation space using two approaches: *first*, by measuring the time derivative of angular displacement (angular velocity) and *second*, by estimating the Hurst exponent (H) [180]. The time derivative of angular displacement determines the change in the angle with time; the smaller the change, the smoother the trajectory and vice-versa. It is approximated by:

$$\dot{\theta}_t = \theta_t - \theta_{t-1}$$

where $t = 2, 3, \cdots, n$ and n = total number of frames in the video. Fig. 7.2(a) plots the time derivatives of angular displacement of the emotion points during each conversation in the testing set. Smaller values imply smooth emotion trajectories.

The Hurst exponent is a statistical measure of persistence and predictability of a time series, calculated by rescaled range $\left(\frac{R_t}{S_t}\right)$ analysis, where R_t is the rescaled

range and S_t is the standard deviation of the time series. The calculation of the rescaled range R_t of the time series will be described shortly. The greater the value of H ($0.5 < H < 1$), the smoother the time series, $H = 0.5$ means a random time series. Fig. 7.2(b) plots the the Hurst exponent as a measure of smoothness of emotion trajectories in the activation-evaluation space for each conversation in the testing set. We estimated H for the time series representing the size of 'change' between pairs of consecutive points in the space as a function of valence and activation. Suppose X_t denotes the time series where $t = 2, 3, \cdots, n$ and n = total number of frames in the video. The size of 'change' for each frame is calculated as the Euclidean distance between the time derivative of valence ($\dot{V}_t$) and the time derivative of activation ($\dot{A}_t$) in the activation-evaluation space:

$$\dot{V}_t = V_t - V_{t-1}$$
$$\dot{A}_t = A_t - A_{t-1}$$

where the t index represents the t^{th} element of the time series.

$$\text{size of change} = \sqrt{\dot{V}_t^{\,2} + \dot{A}_t^{\,2}} \tag{7.1}$$

The rescaled range R_t of time series (X_t) is calculated by:

1. Calculate the mean-centred time series $Y_t = X_t - \mu$, where μ is the mean of the time series.

2. Calculate the cumulative sum of Y_t,

$$Z_t = \sum_{i=1}^{t} Y_i$$

3. Calculate the rescaled range R_t of time series,

$$R_t = \max(Z_1, Z_2, \cdots, Z_t) - \min(Z_1, Z_2, \cdots, Z_t)$$

S_t is the standard deviation of the time series. The ratio $\frac{R_t}{S_t}$ scales as a power law with time so that:

$$H = \frac{\log\left(\left(\frac{R}{S}\right)_t\right) - c}{\log(t)}$$

where c is a constant and the slope of the regression line (R/S versus t in log-log axes) approximates the Hurst exponent. Figure 7.3 shows the linear regression model fitted to the R/S analysis of all emotion trajectories in the testing set.

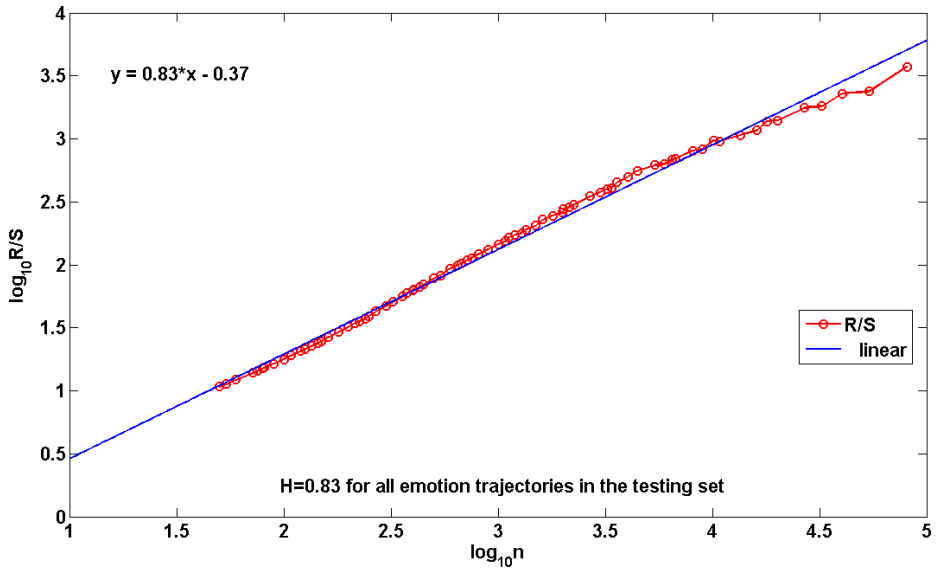


Figure 7.3: A linear regression model fitted to R/S analysis for all emotion trajectories in the testing set.

Fig. 7.4 plots the size of 'change' between two consecutive emotion points in the activation-evaluation space for each conversation in the testing set separately. The histogram shows that the size of 'change' (which is the motion within 120^{th} of a second) is mostly very small. However, in a few cases the size of change is bigger. In order to find the reason behind these intensity jumps (those beyond the first standard deviation), we monitored those paths of trajectories and compared them with the original videos in the dataset. We noticed that the bigger size of change in the trajectories are false positives due to closing the eyes. We had tried to avoid

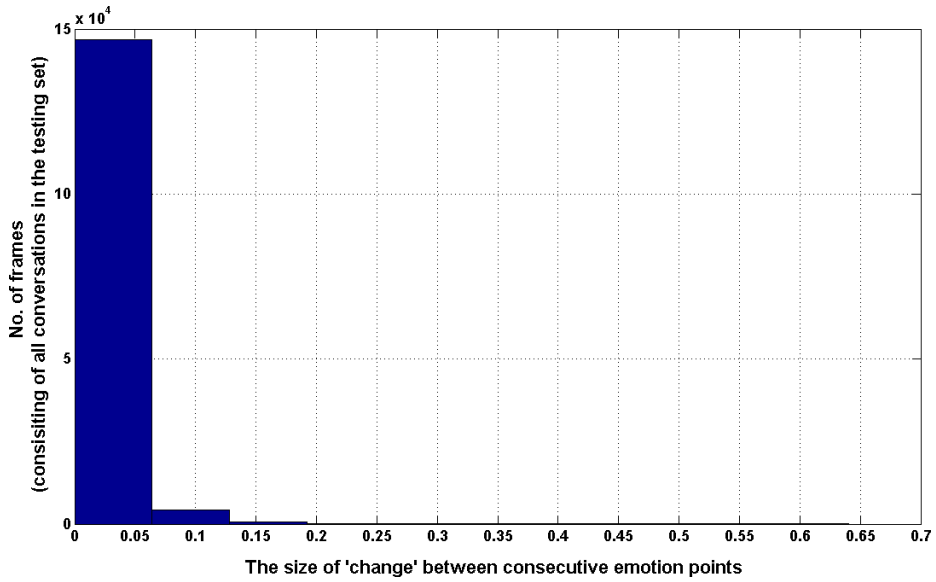


Figure 7.4: The size of the 'change' between two consecutive emotion points in the activation-evaluation space, for all conversations in the testing set. The maximum possible motion is 2, since the circle is radius 1.

this by removing the eyelid markers, but still the closing of eyes is captured by the muscles around the eyes, especially those near the eyebrows. Fig. 7.5 shows some of the false positives during transition from excitement to frustration. Fig. 7.1 also shows that false positives arise due to closing the eyes. As lowering eyelids/eyebrows is one of the expressions of sadness, the outliers in the emotion space lies in the direction of sadness. It should be noted that the size of 'change' refers to the size of displacement between the two consecutive points in the activation-evaluations space, not the displacement of markers on the face.

Hypotheses 2 and 3: Transitions between uncorrelated or negatively correlated emotions need to pass through the neutral state, while transitions between positively correlated emotions do not. According to differential emotion theory, each discrete emotion is related to certain other discrete emotions with a

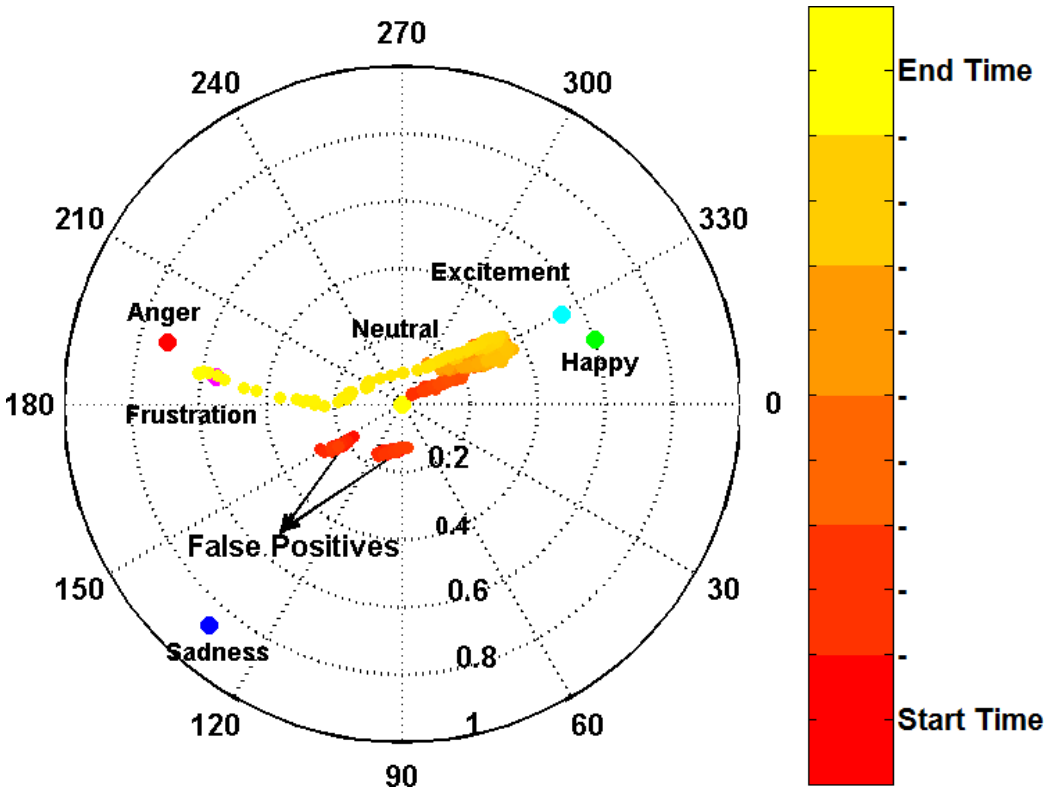
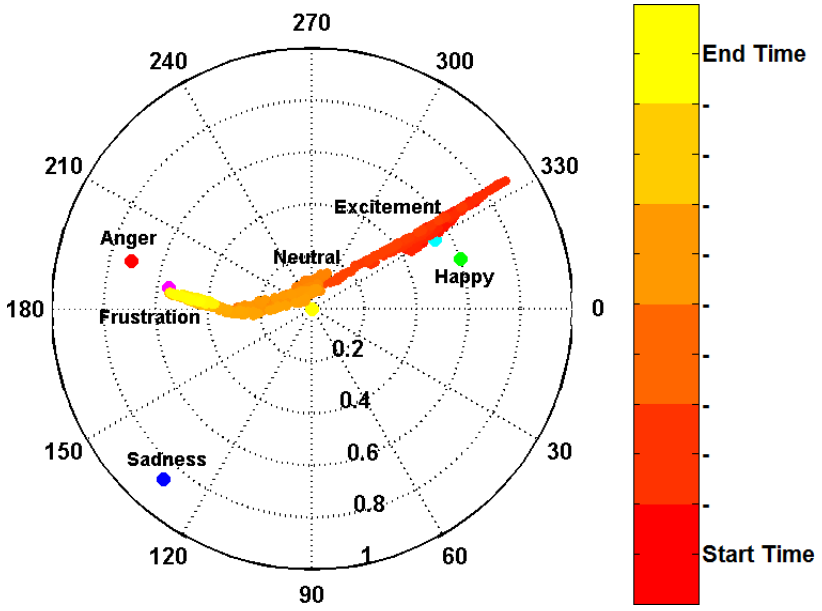


Figure 7.5: The false positives appear during transition from neutral to excited. The movement of emotions through time is represented by changing colour spectrum from dark/red (start) to light/yellow (end).

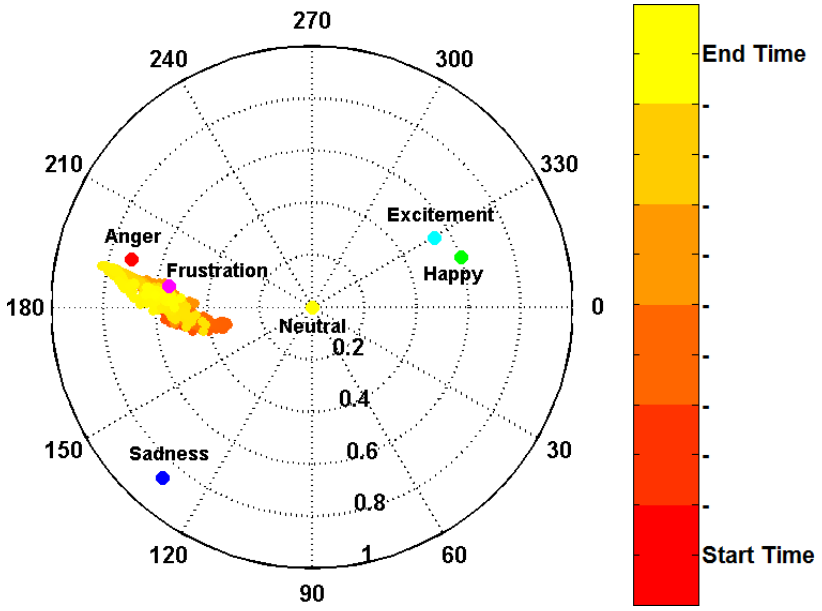
distinct pattern [23]. In the activation-evaluation space, emotions lie along particular angles on the basis of their similarity measures.

While looking at continuous trajectories in the activation-evaluation space (e.g., Fig. 7.1), we noticed some patterns of emotion transitions. We fitted regression lines to the trajectories among six emotions (neutral, happiness, excitement, anger, frustration, and sadness). There are 15 possible symmetric paths among the six emotions, but the regression analysis shows that actually only nine paths are represented for all possible transitions among these emotions. For instance, the *smiling* expression (getting happy) shows a linear relationship of valence and activation between neutral and happiness. The onset of smiling occurs at neutral, reaches apex at some intensity of happiness and returns to neutral for offset. In the case of *laughing*, which is

another expression of happiness, the same pattern repeats several times depends on its duration and intensity.



(a) Continuous transition between excitement and frustration, which are negatively correlated emotions



(b) Continuous transition between anger and frustration, which are positively correlated emotions

Figure 7.6: Transitions between negatively correlated emotions tend to pass through the neutral state, while transitions between positively correlated emotions do not.

By observing the trajectories between two uncorrelated and negatively correlated emotional states, we found that the transition between these states tends to pass through the neutral state. The intensity of the current emotion must decrease to neutral (shown as linear motion along a radial line) before the intensity of the next emotion increases. Fig. 7.6(a) shows a trajectory followed by the transition from excitement to frustration. However, the positively correlated emotions (such as anger and frustration, as well as happiness and excitement) may move from one state to another with slight change in intensity and angle simultaneously, as shown in Fig. 7.6(b). These findings are also supported by the mechanical properties of facial dermal tissues [207]. Under low stress (transition between two positively correlated emotions), dermal tissue applies low resistance to stretch as the collagen fibres uncoil in the direction of the strain. However, under high stress conditions (transition between two negatively or uncorrelated emotions), the elastin fibres behave like elastic springs to return the collagen fibres to their original no-stress condition. According to these properties, to express a very different emotion the facial muscles have to pass through a 'no-stress' condition.

Table 7.1 shows a comparison among the coefficient of determination (R^2) of linear, quadratic, and cubic polynomial regression models fitted to each of the fifteen symmetric paths of emotion transitions into the activation-evaluation space. The table is divided into three sections: the top one shows transitions from neutral to each of the five emotions, i.e., between neutral and anger, neutral and frustration, neutral and happiness, neutral and excitement, and neutral and sadness. For the paths from neutral to any of the five emotions, there is no significant difference among the R^2 values of linear, quadratic and cubic regression ($p > 0.05$), which implies that a linear model may be used to fit these paths.

E1		**E2**	**Linear** (R^2)	**Quadratic** (R^2)	**Cubic** (R^2)
Neutral					
Neu		Ang	0.9131	0.9297	0.9139
Neu		Fru	0.9642	0.9854	0.9924
Neu		Hap	0.9223	0.9323	0.9418
Neu		Exc	0.9374	0.9383	0.9388
Neu		Sad	0.902	0.9144	0.9164
Negatively Correlated Emotions					
Ang		Hap	0.2897	0.6803	0.6856
Ang	Neu		0.9595	0.9632	0.9673
	Neu	Hap	0.9639	0.9967	0.997
Ang		Exc	0.0154	0.6147	0.6897
Ang	Neu		0.9656	0.9839	0.984
	Neu	Exc	0.8946	0.8969	0.8995
Fru		Hap	0.2399	0.4939	0.4981
Fru	Neu		0.9768	0.9811	0.9839
	Neu	Hap	0.9296	0.9362	0.9684
Fru		Exc	0.0504	0.4826	0.6628
Fru	Neu		0.7357	0.7733	0.7894
	Neu	Exc	0.8601	0.897	0.8983
Sad		Hap	0.0604	0.7529	0.7902
Sad	Neu		0.91	0.9138	0.9142
	Neu	Hap	0.9535	0.9545	0.9548
Sad		Exc	0.0564	0.6603	0.6973
Sad	Neu		0.7915	0.8952	0.8977
	Neu	Exc	0.887	0.9088	0.9092
Positively Correlated Emotions					
Ang		Fru	0.6901	0.8735	0.8798
Hap		Exc	0.6097	0.8481	0.8736
Ang		Sad	0.8562	0.996	0.9858
Fru		Sad	0.8177	0.9947	0.9674

Table 7.1: Coefficient of determination (R^2) of linear, quadratic, and cubic polynomial regression models fitted to the fifteen symmetric paths of emotion transitions into the activation-evaluation space. The emotion categories are denoted as *Neu* for Neutral, *Ang* for Anger, *Fru* for Frustration, *Hap* for Happiness, *Exc* for Excitement, and *Sad* for Sadness

The middle section of Table 7.1 shows all paths of transitions among negatively correlated emotions, i.e., between anger and happiness, anger and excitement, frustration and happiness, frustration and excitement, sadness and happiness, and sadness

and excitement. The R^2 values of linear, quadratic and cubic regression between two negatively correlated emotions show that one linear model does not fit well to these paths, while the quadratic and cubic models fit relatively better than one linear model. However, it is clear that the two linear models fit significantly better than the one linear, one quadratic, or one cubic model separately. The table shows the R^2 values of both paths of transition between two negatively correlated emotions, e.g., one directly from anger to happiness and the other from anger to neutral and then from neutral to happiness. This suggests that the negatively correlated emotions tends to pass through the neutral state while transitioning from one state to another.

The last section of Table 7.1 shows all transitions among positively correlated emotions, i.e., between anger and frustration, happiness and excitement, anger and sadness, and frustration and sadness. For the path between these neighbouring emotions, the quadratic regression is better than the linear regression. Moreover, there is no significant difference between quadratic and cubic regressions, which implies that the quadratic model may be used to fit the trajectories between the two neighbouring emotions into the activation-evaluation space. This suggests that the positively correlated emotions can move from one state to another with a slight change in intensity and angle simultaneously.

In the activation-evaluation space, the travel along the emotion flows/trajectories is a matter of intensity change and the angle change. As already discussed, the emotion trajectories follow 'common' paths, which in turn suggests that there exists some relationship between the intensity change and angle change through time. In order to analyse this relationship, we plot the polar coordinates ($\dot{r}$: changing intensity, $\dot{\theta}$: changing emotion) of continuous points in the space during emotion transitions. Fig. 7.7 consists of four subplots, the first and second subplots show r and θ respectively. In these plots, the three horizontal lines represent the mean (μ) and mean $\pm$

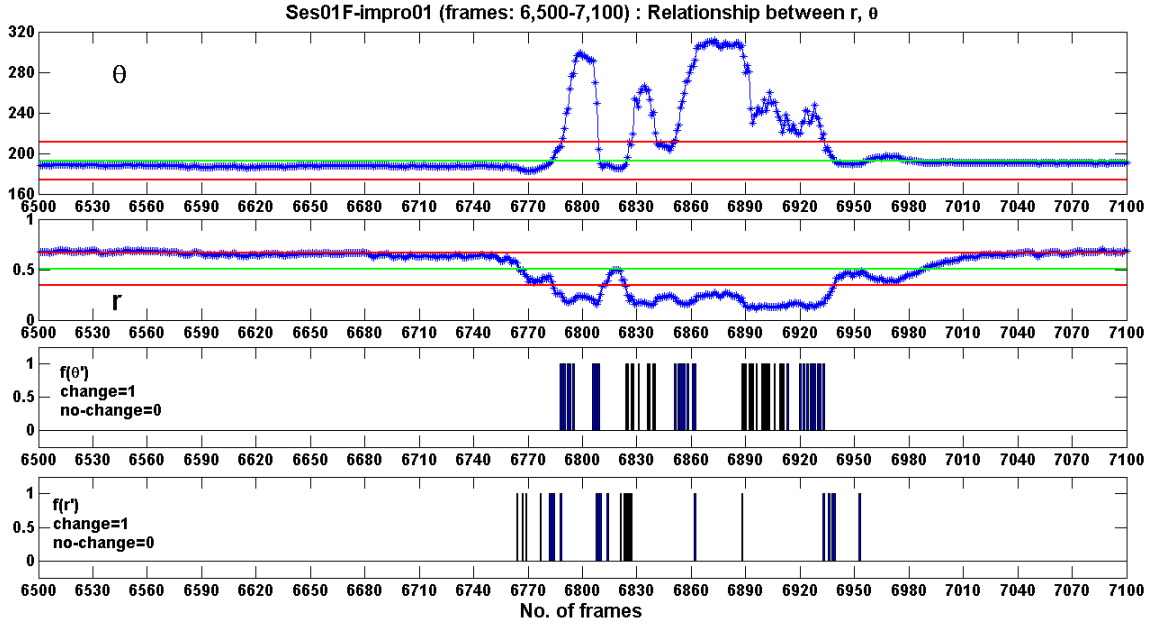


Figure 7.7: Ses01F_impro01, continuous frames during emotion transition from anger to happiness. The temporal relationship between intensity change and angle change can be seen. For explanation, see the text.

1 standard deviation σ. The third and fourth subplots show change versus no-change using a binary plot by applying the following rules to the time derivatives $(\dot{r_t}, \dot{\theta_t})$ of (r, θ) respectively (where the t index represents the t^{th} element of the time series):

$$f(\dot{\theta_t}) = \begin{cases} 0, & \text{if } \mu_{\dot{\theta_t}} - \sigma_{\dot{\theta_t}} < \dot{\theta_t} < \mu_{\dot{\theta_t}} + \sigma_{\dot{\theta_t}} \\ 1, & \text{otherwise} \end{cases}$$

$$f(\dot{r_t}) = \begin{cases} 0, & \text{if } \mu_{\dot{r_t}} - 2\sigma_{\dot{r_t}} < \dot{r_t} < \mu_{\dot{r_t}} + 2\sigma_{\dot{r_t}} \\ 1, & \text{otherwise} \end{cases}$$

Fig. 7.7 shows that there is a relationship between angle change and intensity change such that whenever there is a large 'change' in angle (according to the given rules), the intensity decreases.

7.4 Validation

We used a 4-fold cross-validation method to evaluate the consistency of our models. We randomly chose 10,000 frames of each of the six emotions to form a total set of 60,000 frames. This data includes all frames from the original training set that was used to build the shape models, as well as another 36,000 frames that represented part of a conversation that included emotion changes. For all these frames at least two of the human evaluators agreed on the same label. We set up four different datasets, each of which contained 45,000 frames for training and 15,000 frames for testing. Each of these datasets was used to build new shape models, and then map the emotion trajectories in activation-evaluation space using the method described in Section 6.3. The purpose of this validation is to test the consistency of each model and the reliability of the obtained results. We also tried more than 4 partitions of the total dataset for cross-validation, but the small size of individual datasets caused a lot of variation in the pattern of resulting trajectories.

Fig. 7.8 shows the outputs of the emotion trajectories for the four different testing sets in the activation-evaluation space. It can be seen that the first, third, and fourth show very similar shapes to each other, but that the second one does not have any instances of sadness. However, the match is still very good.

7.5 Conclusion

In this chapter, we have presented an analysis of emotion trajectories in the activation-evaluation space based on shape models of facial points. On the basis of trajectory-level analysis, we evaluated some hypotheses related to the smoothness of emotion trajectories, and the 'common' paths among emotions based on their correlation.

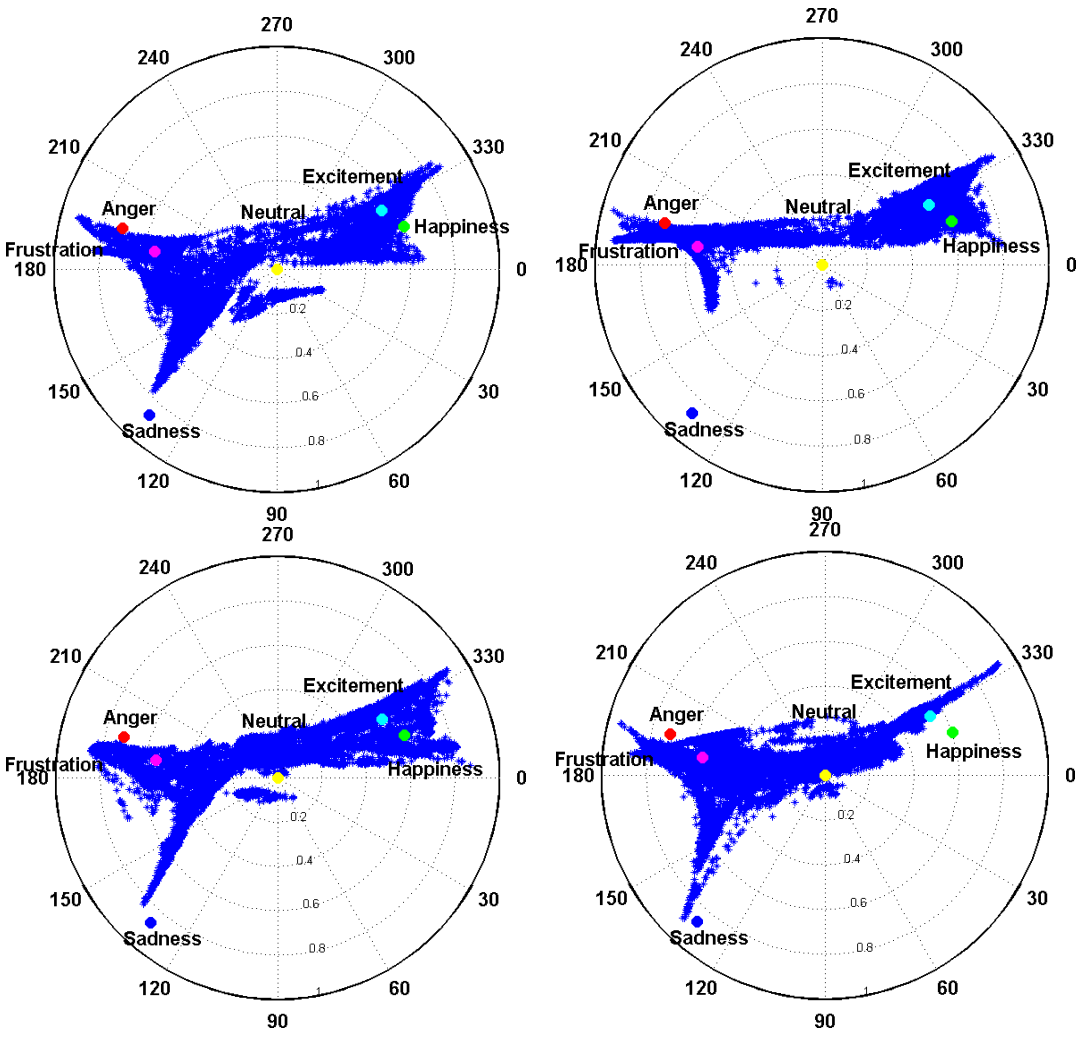


Figure 7.8: Outputs of the emotion trajectories for the four different testing sets in the activation-evaluation space.

By measuring the size of 'change' between two consecutive frames, we found that the emotions move in a continuous flow, which implies that there are no sudden jumps within the trajectories. Further, we measured the smoothness of continuous emotion trajectories on the basis of the time derivative of angular displacement and the estimated Hurst exponent, which suggests that the emotion trajectories are smooth and persistent with time.

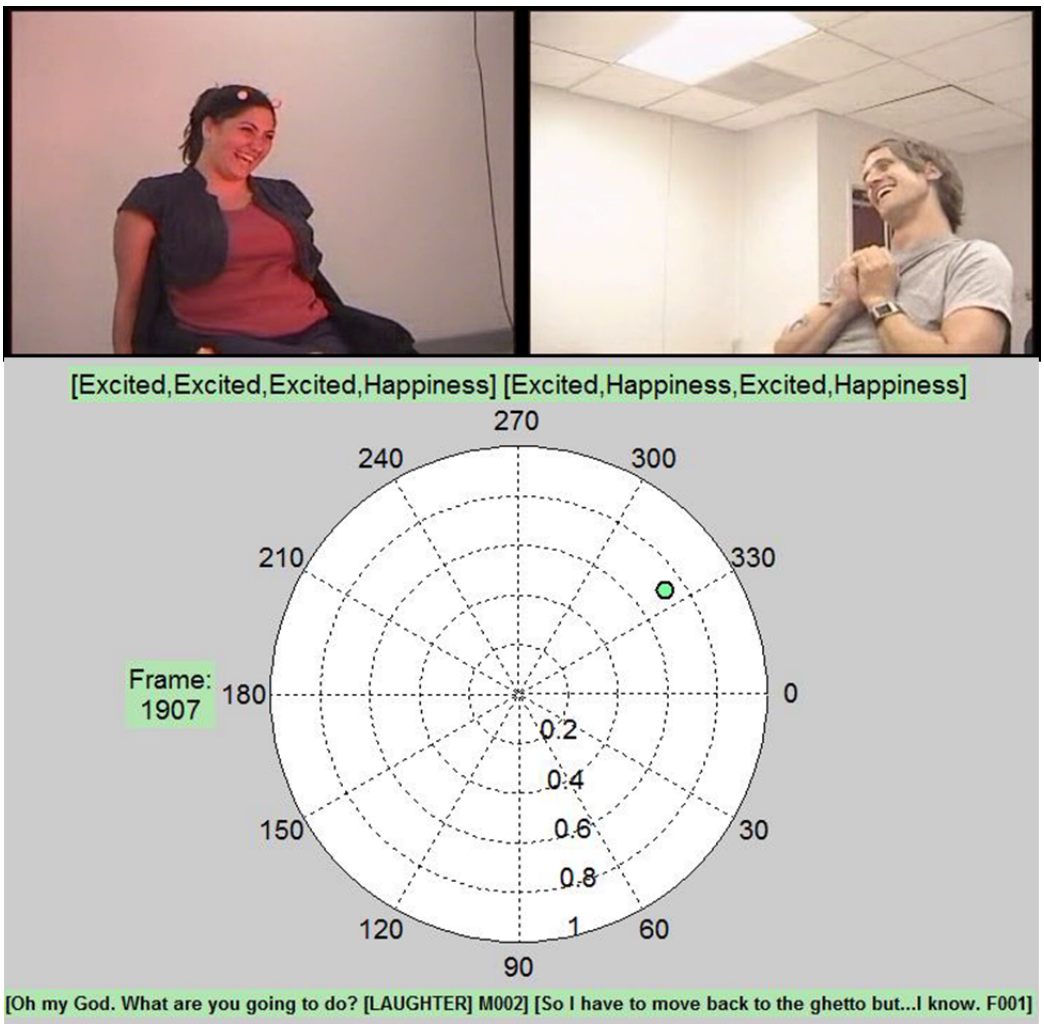


Figure 7.9: A synchronised frame-by-frame comparison of videos from the dataset and the corresponding mapped emotions into the activation-evaluation space. The snapshot shows the *laughing* expression during a happy conversation mapped as *excitement* of high intensity on the space. The female actor was being recorded.

By visualising the emotion trajectories, we observed that there are 15 possible symmetric paths among the six emotions in the activation-evaluation space. To test it, we fitted regression lines to the trajectories and found that there are actually only 9 symmetric paths to travel among these six emotions. We showed that a linear model fits well to the trajectories between neutral and any of the other five emotions. The trajectories between uncorrelated and negatively correlated emotions cannot be

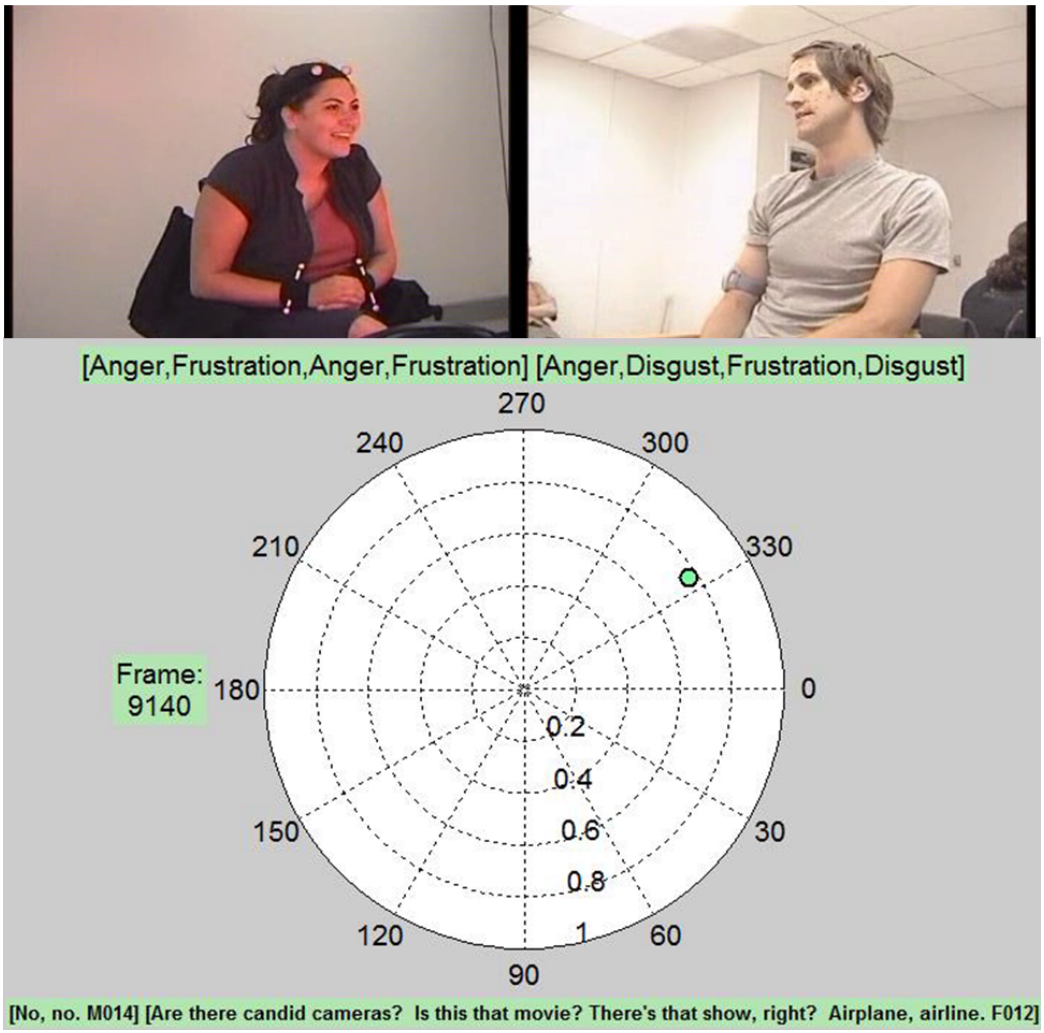


Figure 7.10: The sarcastic expression lead to an abnormal transition towards happiness during an angry conversation. The female actor was being recorded.

fitted with one linear model, however, two linear models are better fitted than one quadratic or cubic regression model. A quadratic regression model fits well to the trajectories between positively correlated emotions.

By analysing the relationship between the change in angle and change in intensity, we may conclude that the transition between negatively correlated or uncorrelated emotions causes a decrease in intensity, while the transition between positively correlated emotions may occur with a slight change in intensity and angle simultaneously.

The frame-by-frame comparison between the emotion trajectories and that of the original videos from the dataset helps to find reasons of some 'abnormal' trajectories. Fig. 7.9 shows a snapshot of a synchronised side-by-side comparison of the original videos from the dataset and the corresponding mapped emotions on the activation-evaluation space during a spontaneous conversation. The ground truth emotion labels by the human observers are displayed at the top and the transcriptions are displayed at the bottom to give an idea about the interaction context through time. We noticed some interesting trajectories which might help to understand natural emotional behaviour during face-to-face conversations. For example, during an angry conversation an abnormal transition towards happiness was caused by a *sarcastic* gesture (as shown in Fig. 7.10).

The presented analysis might be used and extended in several directions, such as examining the 'abnormal' paths of emotions, which might give some cues about underlying deception, or some illness. The mapping of continuous trajectories to the activation-evaluation space might be a useful tool to build emotional conversation agents displaying realistic emotions and going through smooth emotion transitions.

Chapter 8

Conclusions

This final chapter provides an overview of the research and highlights the significant findings of this thesis. The chapter ends with a discussion of opportunities for continued research.

8.1 Research Overview and Significant Findings of the Thesis

This study set out to develop a computational system that can map facial points to some appropriate space and study the trajectories followed by them while a person experiences a set of emotional states through time.

Most research in this area has focused on developing computational systems for the recognition of discrete basic emotion categories. Many of these systems are based on static facial images of posed facial expressions recorded in strict laboratory settings (see Chapter 3). However, research in psychology has shown that our natural emotions are not basic; we usually demonstrate more than one emotional state at a time. Mixtures or blends of basic emotions are referred to as complex emotions (see Chapter

2). Although there are names for many complex emotions, they are often subjective. In addition, the number of them makes categorising them very hard. This is particularly true when the emotions are not posed. Also, most of the work has focused on the final appearance change (e.g., the full-blown emotions) with an exception of some work on the recognition of a few temporal segments (i.e., onset, apex, and offset) of facial expressions rather than dynamic changes in their appearance over time.

This thesis has addressed these problems by focusing on the development of computational methods and techniques dealing with the continuous spontaneous expressions of emotions and their temporal dynamics in order to get insight about the paths followed by emotions while moving from one state to another and the corresponding intensity variations over time. The scope of this thesis is restricted to the use of facial points as a representation of underlying emotional states and mapping these points to the activation-evaluation space, which is the most widely-used representation of emotions in psychological studies. It also includes the study of emotion trajectories in this space to analyse the 'common' paths between emotions and how the intensity varies while moving from one emotional state to another.

Chapter 1 of this thesis began by introducing the term 'affective computing' and some of the potential benefits of having computers recognise and express human emotions. It also presented the problems associated with automatic emotion recognition and analysis of their temporal dynamics, followed by the scope of this thesis. The aims of the research were introduced along with a set of objectives.

Computational study of emotions is an interdisciplinary field using the concepts and theories from psychology and implementing them by using methods from computer vision and machine learning. Chapter 2 sought to build a bridge between these two fields by briefly describing emotions from the perspective of psychology and the fundamental emotion theories necessary for the computational modelling of

emotions. The main concepts described in this chapter include the discrete basic emotions, the psychoevolutionary theory describing complex emotions as mixtures of basic emotions, and the activation-evaluation space that is the most commonly-used representation of emotions in psychological studies.

After getting an insight about emotions from the psychological point of view, the thesis moved on to the discussion of some of the widely-used and state-of-the-art methods of automatic emotion recognition and analysis. Chapter 3 presented an overview of the existing work by describing two major computational approaches to automatic emotion recognition: the direct classification approach, and mapping emotions to emotion space. It also presented some existing techniques for the analysis of temporal dynamics of emotions. All these methods attempted to develop computational systems using concepts and theories from psychological studies, but restricted in some way or another. This chapter pointed out the limitations of each of these systems and explained why the computational study of emotions deserved further study.

Following on, Chapter 4 presented an overview of some of the most widely-used existing datasets for the emotion-related computational studies. The criteria for the selection of a suitable dataset were listed, followed by a brief overview of data annotation tools used for the labelling of discrete as well as continuous emotions. On the basis of the presented survey, this chapter described our selection of the Interactive Emotion Motion Capture (IEMOCAP) dataset along with the associated benefits and limitations. The purpose of this chapter is to explain the selected dataset and how the data was preprocessed to be used further in the modelling and testing of our proposed methods.

The main findings of this thesis were described in Chapters 5, 6, and 7. Chapter 5 introduced a direct classification approach to map a set of facial points to the

discrete basic emotion categories. Based on the research findings that some expressions (mainly negative emotions, such as anger and sadness) are better recognized from muscle activity in the upper half of the face, while others (including positive emotions, such as happiness) from the lower half of the face, this thesis attempted to improve the classification of basic emotions by developing separate shape models of full, upper, and lower parts of the face separately. This method was based on Principal Component Analysis (PCA), which is a statistical machine learning technique useful for analysing sets of datapoints in high dimensional spaces. Three separate shape models were created using PCA and the classification was done using the Mahalanobis Distance (see Section 5.2). It gave us a set of distances of each frame/set of facial points to each of the six basic emotions.

This chapter presented a detailed analysis of these shape models and observed that the first principal component (PC) of the full and lower parts of the face was primarily correlated with talking and therefore not directly connected with emotion recognition (see Section 5.3). For the task of emotion recognition, talking is a confusion factor so we discarded the first PC, but this might be used for the task of speaker recognition if combined with the technique of human face recognition [212]. The first PC can be used to identify that 'someone is talking' and by using face recognition technique it can be identified who is talking. Automatic speaker recognition has many benefits, such as, automatically focusing a camera on the speaker in a gathering, or separating the speech produced by different speakers from each other in a video [16].

Experiments with the proposed joint face model were reported in Section 5.5. In the first experiment, the performance accuracy of this model was compared to that obtained by using some of the widely-used methods of emotion recognition, including a rule-based emotion classifier (Section 5.4.1) and Support Vector Machines (SVM) (Section 5.4.2). It was also tested against the separate shape models of full, upper, and

lower face alone. The experimental results demonstrated that there was no significant difference between the mean accuracy of the proposed model and the SVM classifier. However, the SVM classifier was experimentally shown to be less reliable than the joint face model. The joint face model outperforms separate shape models alone and the rule-based emotion classifier. The robustness of the joint face model was tested in the second experiment (Section 5.5.1), in which some proportion of the training data was manually mislabelled and the first experiment reran. The experimental results demonstrated that the separate shape models as well as the joint face model was resistant until 25% of the training data was mislabelled, while the SVM classifier did not show consistent performance.

The graphs in Figure 5.14 presented the plots of Mahalanobis distance of some of the testing frames in sequence with time to the different clusters of basic emotions. Section 5.6 analysed these graphs and showed that the classification of facial points to discrete basic emotions is not an appropriate way to describe spontaneous emotions. Motivated by these observations, the next chapter presented a method of representing complex emotions as blends/mixtures of basic emotions instead of discrete categories.

Chapter 6 presented a discussion of circular data analysis and described why linear statistical models are not suitable to describe circular data. This chapter introduced a method of representing complex emotions in the activation-evaluation space by using a degree of match of each frame/set of facial points to each basic emotion given by the shape models described in Chapter 5. The activation-evaluation space describes a disk of potential emotions, representing them in terms of polarity and their similarity to each other. In the past, quite a lot of methods attempted to map emotions to the activation-evaluation space (see Section 1.3), but none of them considered that a variable describing emotional data on this space is a circular variable.

The method described in Chapter 6 is based on a statistical modelling technique based on a mixture of von Mises probability distributions, which is a circular analogue of the normal distribution. It enables us to represent the distributions of six basic emotions and their mixtures on the activation-evaluation space, in order to interpolate complex emotions.

Experiments with this method were reported in Section 6.4. The first experiment compared the locations based on the direction and intensity of the basic emotions on the activation-evaluation space estimated by the model to that listed in the psychological research conducted by Whissell [221]. The locations of all basic emotions matched reasonably except that of the neutral state. This was because human evaluators misinterpreted neutral as sadness and assigned a very low activation value, as was discussed in Section 6.4.

In the second experiment, the mean direction and intensity of individual utterances as well as full conversations were compared to those estimated by using the ground truth values of valence and activation using the method described in Section 6.3.1. The circular sample correlation coefficient was used to find the correlation between the mean direction estimated by the model and that of the ground truth. The intensity values were compared using the pairwise t-test, showing significant association. In the third experiment, the goodness of fit of two samples (ground truth values and those estimated by the model) was proved using Kuiper's test (see Section 6.4).

The method described in Chapter 6 enabled us to represent the emotions in the activation-evaluation space. By observing the paths of emotional expressions on this space through time, we came across specific sequences of patterns: an example was shown in Fig. 7.1. This visual pattern and the effect of mechanical properties of human skin on facial expressions motivate us to analyse these emotion trajectories.

Chapter 7 sought to test the hypotheses made in Section 1.3.2 about the paths of emotions representations in the activation-evaluation space. This was done by considering the locations of facial points on the space as a time series.

The processes used to test the hypotheses were described in Section 7.3. To test the first hypothesis, the time derivative of angular displacement was calculated and the Hurst exponent (H) was estimated using the time series of facial points locations on the activation-evaluation space. The small values of angular displacement and the large values of H proved that the emotions follow smooth trajectories in the space. In the second test, regression lines were fitted to the trajectories between emotions, which found that the paths between uncorrelated or negatively correlated emotions need to pass through the neutral state, but this is not necessary for positively correlated emotions. This showed that the second and third hypotheses were true.

The chapter also presented an experiment which demonstrated that there is a relationship between angle change and intensity change in such a way that whenever there is a large change in the direction of an emotion path, the intensity of that emotion decreases. Finally, all the processes demonstrated above were tested using 4-fold cross-validation to evaluate the consistency and reliability of the obtained results as described in Section 7.4. Four different datasets were created using the total set of 60,000 frames (10,000 frames of each of the six emotions), out of which 45,000 frames were used for training and 15,000 frames were used for testing. The whole process, starting from building the shape models to mapping facial points to the activation-evaluation space using the mixture model, was repeated four times. The outputs of resulting emotion trajectories showed a very good match.

All the three objectives listed in Section 1.3.2 were fulfilled in this thesis. The first objective, which was to map the set of facial points to the weighted basic emotions

was fulfilled in Chapter 4 and 5. A set of 5 basic emotions (anger, frustration, happiness, excitement, sadness) and the neutral state were selected based on the reasons described in Section 4.7. The facial points were modelled using a combination of separate shape models of full, upper, and lower parts of the face based on Principal Component Analysis (PCA) (Section 5.2). The performance accuracy and robustness of the model was tested by comparing with other widely-used methods and the ground truth data based on the majority categorical labells assigned by three human evaluators (Section 5.5). The results showed considerable improvement in the accuracy and robustness of basic emotion recognition using the proposed joint face model. The joint face model gave a set of distances (the reciprocal of these distances were used as *weights*) of each frame/set of facial points to each of the six emotions using the Mahalanobis distance (Section 5.2.2). These distances were used to describe complex emotions as a mixture/blend of basic emotions.

The second objective, which was to represent emotions in some space was fulfilled in Chapter 6. Based on the psychological literature, the activation-evaluation space was selected as a suitable space to represent emotions (Section 2.5). A method based on von Mises mixture model was developed that is capable of mapping basic emotions into this space (Section 6.3). Using the set of distances obtained by using the shape models, this method can represent each frame/set of facial points as a blend/mixture of basic emotions in the activation-evaluation space (Section 6.3). The accuracy of this method was evaluated by comparing the estimated facial points locations (direction and intensity of emotional expression) with those obtained from the ground truth data (Section 6.4). The results showed that the proposed mixture model fits the data well.

The third objective, which was to analyse the paths of emotion representations on the space by considering it as a time series, was fulfilled in Chapter 7. All the three

hypotheses about the emotion paths representations (listed in Section 7.2) were tested in Section 7.3. The results were evaluated by cross-validation using four different sets of training and testing sets in order to test the consistency of our models. The validation demonstrated an overall good match of the obtained sequence of emotion trajectories on the activation-evaluation space (Section 7.4).

8.2 Future Research

The methods proposed in this thesis have been shown to be promising to address the problems of emotion recognition, representation, and analysis. However, there are other interesting properties that have not been explored, which we believe merit further study. In this section, we outline some directions for future research.

The first suggestion is about how to implement the proposed system in practical real-life applications by using full facial images instead of marker point locations. The second and third suggestions are about how to include additional information that could improve automatic emotion recognition. The fourth suggestion is about observing the time-offset of emotion activation and recovery after the emotional stimuli to maintain emotional well-being. The fifth suggestion is about preserving old knowledge while gaining new information about the paths of emotion transitions in order to learn the temporal structure of one's emotional life.

8.2.1 Using full facial images instead of marker locations

In this thesis, we have used the markers on the face to capture the facial changes caused by underlying emotional states. This was a simplification to avoid preprocessing steps of face detection and facial feature extraction from videos. However, it is obviously not appropriate in real-life practical applications. In order to use the

proposed methods in some practical or commercial application, the system needs to automatically detect and track faces and the useful facial features for emotion recognition.

We have worked on the development of a reliable face detection system which is capable of localizing and extracting the face region from the rest of the image/video frame [99]. The system is based on a combination of Haar classifiers [216] and skin colour detection methods [133], and is more reliable than using any of them separately. We plan to improve this work to be able to track faces in real-time and to track the facial points automatically without using markers.

8.2.2 Emotion recognition using multiple modalities

This thesis used the facial points as a representation of emotions, however, human emotional information can be obtained from a broad range of behavioural clues and physiological signals, such as tone of voice, words, head and body gestures, heart rate, skin conductivity, and body temperature. After face, voice is the second most useful expression of emotion to be used in the computer systems, followed by head and hand movements and body gestures. Their occurrences are temporally synchronized, which makes it possible to use them for improving the recognition of emotions [96].

Humans need to interact in all possible situations and develop an adaptive behaviour, but machines can be specialised depending on the needs. To improve the emotion recognition methods described in this thesis, it will be useful to add voice data as well as head movements to get additional information about the actor's emotional states. Another advantage of using multiple clues is that in real-life scenarios, if one clue is unavailable, the system can use the other.

8.2.3 Adding contextual information

In the literature related to automatic emotion recognition, very few works have used the contextual information and those that did generally used different definitions of context. Broadly, for the task of emotion recognition and analysis, the context/external stimuli refers to the knowledge of who the subject is, where she is, when she is being observed, and what her current task is [97].

The computational systems that aim to use contextual information include the following: [150, 158] considered the information about past emotions as context to improve current emotion recognition, [228] used multiple clues for emotion recognition and considered one clue as the context of another. In [80], the authors presented an emotion recognition system in the context of a vehicle for monitoring driver's state of alertness, and [224] demonstrated affect-sensitive tutors in the context of a teaching environment. The system presented in [132] attempted to model one person's influence on the interacting partner's behaviours during a dyadic spoken conversation scenario. During an interaction people adapt to each other and a shared equilibrium is formed between them, therefore, it is crucial to assess the reciprocity of interacting participants to improve the automatic emotion recognition and analysis [34].

Defining a context simplifies the problem by limiting the scope to a specific task. For example, in a vehicle's environment, one might expect a frontal view of the subject's face with limited head and body movements. In this thesis, we have used the IEMOCAP dataset where the actors are engaged in dyadic spoken conversations. It will be useful to observe the effect of one speaker's emotions on the other, in order to find reasons for some of the abrupt emotional changes, that remained undefined without the contextual information. We plan to do a side-by-side comparison of continuous synchronised actors' emotion trajectories observed in two separate activation-evaluation spaces. It will show the effect of one actor's emotions on the

other through time, that may be used to find reasons behind abnormal transitions between the emotional states.

8.2.4 Observing the time-offset of emotion activation and recovery

To maintain emotional well-being, the time of emotion activation and the time taken to recover, particularly from negative emotions, are both crucial factors [154]. Most of the research in this area has been done in the fields of psychology and neuroscience. The majority of it focused on the examination of an emotion activation upon the onset of an emotional stimulus; very few studies have investigated the extent to which emotional responding continues after the stimulus offset [98, 110]. So far, the research in computer science is limited to the detection of onset, apex, and offset of a facial expression, but the computational study of the time-offsets of continuous emotions is still an unexplored research area. Examining the time-course starting from emotion activation to recovery (to a baseline state) may provide crucial information to understand adaptive recovery time after an emotion eliciting stimulus [51].

In this thesis, the analysis of graphs shown in Fig. 7.7 leads us to study the time offset between intensity change and angle change before, during, and after the emotion transitions. It appears that the intensity starts changing a few lags before the angle change, and the change continue to occur a few lags after the angle change (see third and fourth subplots in Fig. 7.7). Based on this observation, we calculated the cross-correlation between the change in intensity and the change in angle. Fig. 8.1 shows the cross-correlation ($\pm$250 lags) between intensity change and angle change for the same emotion transition (frames: 6500-7100) as that in Fig. 7.7. The graph shows quite high correlation at the offset of -40 frames ($\frac{1}{3}^{rd}$ of a second) before the emotion

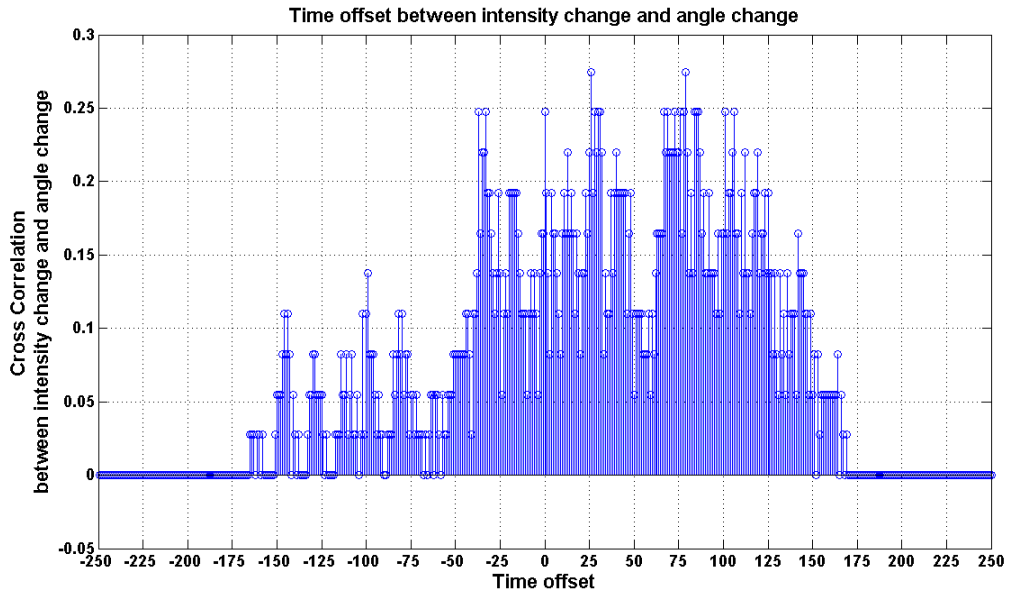


Figure 8.1: Ses01F_impro01, continuous frames (6500-7100) during emotion transitions. Time-offset between intensity change and angle change.

transition and high correlation at the offset of almost 120 frames (1 second) after the emotion transition.

These observations seem promising, but due to time constraints we could not analyse them fully. We plan to investigate these graphs further, to study the time of emotion activation and time of recovery to the baseline state after the emotion transition. It will also be useful to compare different actors to find individual differences in emotion activation and recovery timings.

8.2.5 Life-long learning

Emotional life has a definite temporal structure. Emotions are usually short-lived and fade with time. Spontaneous emotions last for a few seconds, while posed emotions may last for minutes or even hours. Moods describe the emotional states that are underlying and usually last for a long time. Emotions may be effected by moods, while

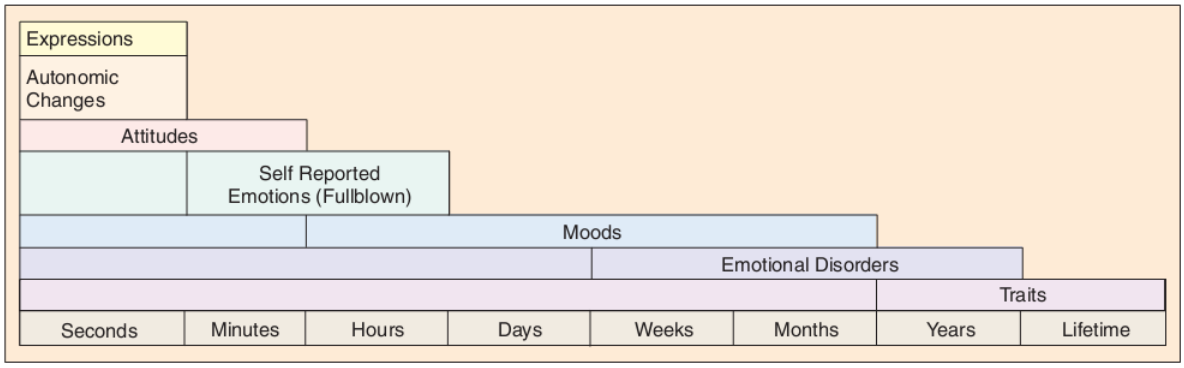


Figure 8.2: The temporal structure of emotional life, figure taken from [44].

moods may be effected by certain emotional disorders like depression or anxiety [44]. Figure 8.2 (taken from [44]) describes the temporal structure of emotional life ranging from seconds to the lifetime.

In this thesis, we focused on the emotional expressions within an utterance lasting for a few seconds (average 4.5 seconds) within full conversations lasting around 5 minutes. We observed several emotion transitions within a conversation, most of which were between the positively correlated emotions. The transition to uncorrelated or negatively correlated emotions were usually observed in special cases, e.g., smiling during a sad conversation or sarcastic expression during an angry conversation. Generally, in all conversations the overall 'mood' of the actor was found to remain the same.

The study of temporal structure of emotional life is a challenging, but very interesting problem. Observing and preserving the emotional information for a long time, even for a life-time, might unveil interesting facts about the temperament of a person, that is, whether that person is briefly angry or has a bad temperament. This information would be useful to deal with patients suffering from emotional disorders such as depression, or the cognitive disabilities such as dementia.

8.3 Conclusions

This thesis has accomplished the task of giving computers the ability to recognise, represent, and analyse emotions using a set of facial points. The study showed that it is possible to classify a set of facial points to basic as well as complex emotions. The shape models developed in this thesis are capable of describing a set of facial points as mixture/blend of basic emotion categories, referred to as complex emotions. The joint face model based on the principal component analysis and the Mahalanobis distance was shown to be more accurate and robust than the widely-used methods of basic emotion recognition. The proposed model is reasonably efficient as the PCA has been performed just once and the classification of unknown frames is done on the basis of their distance from each of the emotion cluster. However, assuming that emotions change slowly through time there is a potential for further improving the efficiency of the proposed model by classifying frames after regular intervals of time rather than including all frames. Since the selected dataset recorded the location of markers at the rate of 120 frames per second, the subsampling of frames should not lose much information.

In this thesis, the facial points have been obtained by using a set of markers on the face using the IEMOCAP dataset, which is not appropriate to be used in practical applications. We plan to automatically extract the facial points to make this system applicable in real-life applications. Among a large number of available audiovisual datasets, the IEMOCAP dataset was selected because it consists of fairly large amount of spontaneous emotion data, annotated in categorical labels as well as in terms of emotion attributes. However, it came up with some limitations such as, assuming the emotional content does not change during an utterance, the labelling was done on the utterance-level instead of the frame-level. Also, there was insufficient data for the complete set of basic emotions which led us to use six (including one neutral,

three basic, and 2 candidate basic emotions) instead of nine emotional states for our analysis.

This study has also investigated how useful it is to represent facial changes through time in an appropriate emotion space. Using the activation-evaluation space for emotion representation, a mixture model has been shown to be a suitable model for characterising emotion data. Using this model, the sets of facial points were mapped continuously through time to the activation-evaluation space. In the training data, the neutral state has been assigned a very low activation value by the human evaluators. This is why it was moved down towards very passive state in the space. To correct this, we deliberately shifted neutral to the centre (close to 0) in the activation-evaluation space. Since, the activation-evaluation space represents a disk of emotions, the variables mapped on this space are circular variables. This thesis has shown that the standard normal probability distribution is inappropriate for describing such data. The von Mises probability distribution has been shown to be suitable for characterising directional emotion data.

Further, this thesis has proposed suitable ways to analyse paths of continuous spontaneous emotions of a person experiencing a set of emotional states. This analysis has shown that the paths between emotions are smooth, and that the emotion transitions follow certain sequence of patterns. Also, the relationship between intensity change and angle change has been investigated during emotion transitions. It showed that whenever there is a large change in angle, the intensity of emotion decreases. We plan to investigate the time-offsets between the angle change and intensity change in order to investigate the timing of emotion activation and recovery after the emotion transition caused by any emotional stimulus.

Appendix A

List of Publications

Parts of this thesis are based on previously published materials as follows:

- A. Hakim, S. R. Marsland, and H. W. Guesgen, "Computational Analysis of Emotion Dynamics," In Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, Switzerland, 2013.

- A. Hakim, S. R. Marsland, and H. W. Guesgen, "Statistical Modelling of Complex Emotions using Mixture of von Mises Distributions," In Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, Switzerland, 2013.

- A. Hakim, S. R. Marsland, and H. W. Guesgen, "A Robust Joint Face Model for Human Emotion Recognition," In Proceedings of the International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 352-357, New Zealand, 2012.

- A. Hakim, S. R. Marsland, H. W. Guesgen, "A Reliable Hybrid Technique for Human Face Detection," In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), pp. 241-244, France, 2010.

Appendix B

Emotion Words from Whissell and Plutchik

	Activ	**Eval**	**Angle**		**Activ**	**Eval**	**Angle**
Accepting	-	-	0	Disgusted	5	3.2	161.3
Adventurous	4.2	5.9	270.7	Disinterested	2.1	2.4	127.3
Affectionate	4.7	5.4	52.3	Disobedient			242.7
Afraid	4.9	3.4	70.3	Displeased			181.5
Aggressive	5.9	2.9	232	Dissatisfied	4.6	2.7	183
Agreeable	4.3	5.2	5	Distrustful	3.8	2.8	185
Amazed	5.9	5.5	152	Eager	5	5.1	311
Ambivalent	3.2	4.2	144.7	Ecstatic	5.2	5.5	286
Amused	4.9	5	321	Elated			311
Angry	4.2	2.7	212	Embarrassed	4.4	3.1	75.3
Annoyed	4.4	2.5	200.6	Empty	3.1	3.8	120.3
Antagonistic	5.3	2.5	220	Enthusiastic	5.1	4.8	313.7
Anticipatory	3.9	4.7	257	Envious	5.3	2	160.3

Anxious	6	2.3	78.3	Exasperated			239.7
Apathetic	3	4.3	90	Expectant			257.3
Apprehensive			83.3	Forlorn			85
Ashamed	3.2	2.3	83.3	Furious	5.6	3.7	221.3
Astonished	5.9	4.7	148	Generous			328
Attentive	5.3	4.3	322.4	Gleeful	5.3	4.8	307
Awed			156.7	Gloomy	2.4	3.2	132.7
Bashful	2	2.7	74.7	Greedy	4.9	3.4	249
Bewildered	3.1	2.3	140.3	Grief			127.3
Bitter	6.6	4	186	Grouchy	4.4	2.9	230
Boastful	3.7	3	257.3	Guilty	4	1.1	102.3
Bored	2.7	3.2	136	Happy	5.3	5.3	323.7
Calm	2.5	5.5	37	Helpless	3.5	2.8	80
Cautious	3.3	4.9	77.7	Hesitant			134
Cheerful	5.2	5	25.7	Hopeful	4.7	5.2	298
Confused	4.8	3	141.3	Hopeless	4	3.1	124.7
Contemptuous	3.8	2.4	192	Hostile	4	1.7	222
Content	4.8	5.5	338.3	Humiliated			84
Contrary	2.9	3.7	184.3	Impatient	3.4	3.2	230.3
Co-operative	3.1	5.1	340.7	Impulsive	3.1	4.8	255
Critical	4.9	2.8	193.7	Indecisive	3.4	2.7	134
Curious	5.2	4.2	261	Indignant			175
Daring	5.3	4.4	260.1	Inquisitive			267.7
Defiant	4.4	2.8	230.7	Interested			315.7

Delighted	4.2	6.4	318.6	Intolerant	3.1	2.7	185
Demanding	5.3	4	244	Irritated	5.5	3.3	202.3
Depressed	4.2	3.1	125.3	Jealous	6.1	3.4	184.7
Despairing	4.1	2	133	Joyful	5.4	6.1	323.4
Disagreeable	5	3.7	176.4	Loathful	3.5	2.9	193
Disappointed	5.2	2.4	136.7	Lonely	3.9	3.3	88.3
Discouraged	4.2	2.9	138	Meek	3	4.3	91
Nervous	5.9	3.1	86	Self-conscious			83.3
Obedient	3.1	4.7	57.7	Self-controlled	4.4	5.5	326.3
Obliging	2.7	3	43.3	Serene	4.3	4.4	12.3
Outraged	4.3	3.2	225.3	Shy			72
Panicky	5.4	3.6	67.7	Sociable	4.8	5.3	296.7
Patient	3.3	3.8	39.7	Sorrowful	4.5	3.1	112.7
Pensive	3.2	5	76.7	Stubborn	4.9	3.1	190.4
Perplexed			142.3	Submissive	3.4	3.1	73
Planful			269.7	Surprised	6.5	5.2	146.7
Pleased	5.3	5.1	328	Suspicious	4.4	3	182.7
Possessive	4.7	2.8	247.7	Sympathetic	3.6	3.2	331.3
Proud	4.7	5.3	262	Terrified	6.3	3.4	75.7
Puzzled	2.6	3.8	138	Timid			65
Quarrelsome	4.6	2.6	229.7	Tolerant			350.7
Ready			329.3	Trusting	3.4	5.2	345.3
Receptive			32.3	Unaffectionate	3.6	2.1	227.3
Reckless			261	Uncertain			139.3

Rebellious	5.2	4	237	Uncooperative			191.7
Rejected	5	2.9	136	Unfriendly	4.3	1.6	188
Remorseful	3.1	2.2	123.3	Unhappy			129
Resentful	5.1	3	176.7	Unreceptive			170
Revolted			181.3	Unsympathetic			165.6
Sad	3.8	2.4	108.5	Vascillating			137.3
Sarcastic	4.8	2.7	235.3	Vengeful			186
Satisfied	4.1	4.9	326.7	Watchful			133.3
Scared			66.7	Wondering	3.3	5.2	249.7
Scornful	5.4	4.9	227	Worried	3.9	2.9	126

Table B.1: Emotion Words from Whissell and Plutchik. List taken from [44]. The first two numerical values represent valence and activation of each emotion word that was found in the study by Whissell [221], and the fourth column represents the corresponding angular location in the activation-evaluation space that was found in the study by Plutchik [174]. The values of valence range from 1 (negative extreme) to 7 (positive extreme), and the values of activation range from 1 (very passive) to 7 (very active).

Appendix C

The effects of varying the second principal component on the female mean upper face

$\mu - \mathbf{3}\sigma$	$\mu - \mathbf{2}\sigma$	$\mu - \mathbf{1}\sigma$	μ	$\mu + \mathbf{1}\sigma$	$\mu + \mathbf{2}\sigma$	$\mu + \mathbf{3}\sigma$
1.2911	1.2754	1.2596	1.2439	1.2281	1.2124	1.1967
30.3317	31.4112	32.4907	33.5702	34.6497	35.7293	36.8088
92.4934	92.2967	92.1000	91.9034	91.7067	91.5100	91.3133
26.1001	26.4326	26.7650	27.0974	27.4299	27.7623	28.0947
23.0419	24.0094	24.9769	25.9444	26.9119	27.8794	28.8469
66.3602	66.0285	65.6969	65.3652	65.0335	64.7019	64.3702
17.6964	17.7697	17.8429	17.9161	17.9894	18.0626	18.1358
26.0558	27.0327	28.0096	28.9865	29.9633	30.9402	31.9171
82.8104	82.6256	82.4408	82.2560	82.0712	81.8864	81.7016
37.6496	37.7765	37.9034	38.0303	38.1572	38.2841	38.4110

36.1744	37.2491	38.3239	39.3987	40.4735	41.5483	42.6231
80.8729	80.6155	80.3581	80.1007	79.8433	79.5859	79.3284
45.7049	46.1155	46.5261	46.9368	47.3474	47.7581	48.1687
34.8941	35.8810	36.8678	37.8547	38.8416	39.8285	40.8154
66.1560	65.5044	64.8528	64.2011	63.5495	62.8979	62.2462
-29.5436	-29.9953	-30.4470	-30.8987	-31.3504	-31.8021	-32.2538
22.3043	23.4113	24.5184	25.6254	26.7325	27.8395	28.9466
68.4880	67.9749	67.4618	66.9487	66.4357	65.9226	65.4095
-18.9518	-19.0602	-19.1686	-19.2770	-19.3854	-19.4938	-19.6022
24.9760	25.9196	26.8631	27.8067	28.7502	29.6938	30.6373
83.1842	82.9360	82.6878	82.4396	82.1914	81.9431	81.6949
-39.4181	-39.5705	-39.7229	-39.8752	-40.0276	-40.1799	-40.3323
35.6584	36.5432	37.4280	38.3128	39.1976	40.0823	40.9671
83.9989	83.4420	82.8850	82.3281	81.7712	81.2143	80.6574
-46.4444	-46.9578	-47.4711	-47.9845	-48.4979	-49.0113	-49.5246
33.5194	34.2373	34.9552	35.6731	36.3910	37.1089	37.8268
68.0680	67.2436	66.4192	65.5947	64.7703	63.9459	63.1214
4.9562	5.1091	5.2619	5.4148	5.5676	5.7205	5.8733
17.1577	17.8481	18.5386	19.2291	19.9195	20.6100	21.3005
59.9792	59.5912	59.2032	58.8153	58.4273	58.0393	57.6513
28.8017	29.1530	29.5042	29.8554	30.2067	30.5579	30.9091
25.0250	25.4442	25.8633	26.2825	26.7017	27.1209	27.5400
51.9217	51.5452	51.1686	50.7921	50.4156	50.0390	49.6625
46.1994	46.6918	47.1841	47.6764	48.1687	48.6611	49.1534
30.8855	31.3227	31.7599	32.1971	32.6344	33.0716	33.5088

53.9253	53.3482	52.7712	52.1941	51.6170	51.0399	50.4629
60.4394	60.7578	61.0761	61.3945	61.7128	62.0312	62.3496
49.1317	49.5916	50.0515	50.5114	50.9713	51.4312	51.8911
46.8489	46.1190	45.3892	44.6594	43.9296	43.1998	42.4700
-7.4811	-7.6252	-7.7693	-7.9134	-8.0575	-8.2016	-8.3457
16.9280	17.6723	18.4166	19.1609	19.9052	20.6496	21.3939
59.3786	59.0162	58.6537	58.2913	57.9289	57.5664	57.2040
-28.4334	-28.6494	-28.8654	-29.0814	-29.2974	-29.5135	-29.7295
22.5947	23.1210	23.6473	24.1735	24.6998	25.2261	25.7523
53.4204	52.9916	52.5629	52.1341	51.7053	51.2765	50.8477
-44.1531	-44.5682	-44.9833	-45.3984	-45.8135	-46.2286	-46.6437
27.3205	27.5758	27.8311	28.0864	28.3418	28.5971	28.8524
56.2417	55.6199	54.9982	54.3764	53.7546	53.1329	52.5111
-61.2547	-61.3401	-61.4255	-61.5109	-61.5963	-61.6817	-61.7671
43.5666	25.4442	43.7889	43.9001	44.0112	44.1224	44.2336
51.7657	50.7500	49.7342	48.7184	47.7026	46.6868	45.6710

Table C.1: The numerical values demonstrating the direction of 3D movement of 17 marker points effected by varying the second PC between $\pm 3\sigma$ on the female mean upper face. μ denotes the mean face and σ denotes the standard deviation. Considering the mean face as a reference, the values identify the upward movement of forehead and eyebrows marker points by varying PC2 from -1σ to -3σ and the downward movement of forehead and eyebrows marker points by varying PC2 from $+1\sigma$ to $+3\sigma$.

Appendix D

Glossary

- **activation-evaluation space** represents emotions based on their activation and valence/evaluation. Activation refers to how motivated a person is during an emotional state. Evaluation refers to how positive or negative the emotion is.

- **affective computing** also known as emotive computing is the computing that relates to, arises from, or deliberately influences emotions.

- **circular data** also known as directional data refers to the data containing observations on a circle of unit radius.

- **continuous emotions** refer to the uninterrupted sequence of emotions which are dynamic in terms of the changing facial patterns, the speed of onset, apex and offset movements, their intensity and duration.

- **emotion dynamics** refer to the continuously varying properties of emotions including intensity, flow, persistence with time, and their relationships with other emotions.

- **emotion trajectories** are the paths followed by emotions into some emotion-space while moving from one state to another through time.

- **end-point emotions** refer to the emotions at the two ends of an emotion trajectory. For example, for a trajectory between anger and happiness, 'anger' and 'happiness' are the end-point emotions.

- **negatively correlated emotions** correspond to the emotions which are 180° apart in the activation-evaluation space.

- **neutral state** refers to a 'no emotion' state, or the presence of any emotion with very low (unnoticeable) intensity.

- **positively correlated emotions** are those that lie close to each other (less than 90° divergence) in the activation-evaluation space.

- **uncorrelated emotions** correspond to the emotions that lie at 90° divergence in the activation-evaluation space.

References

[1] S. Abrilian, L. Devillers, S. Buisine, and J.-C. Martin. EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International*, 2005.

[2] I. Albrecht, M. Schröder, J. Haber, and H.-P. Seidel. Mixed feelings: Expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*, 8(4):201–212, 2005.

[3] M. B. Arnold. *Emotion and Personality*. Columbia University Press, 1960.

[4] S. Bahn, J. Han, and C. Chung. Facial expression database for mapping facial expression onto internal state. In *Emotion Conference of Korea*, pages 215–219, 1997.

[5] A. Bansal, S. Chaudhary, and S. D. Roy. A novel LDA and HMM-based technique for emotion recognition from facial expressions. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, pages 19–26. Springer, 2013.

[6] T. Bänziger and K. R. Scherer. Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus. *Blueprint for Affective Computing: A Sourcebook*, pages 271–294, 2010.

[7] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22, 2001.

[8] L. F. Barrett and E. Bliss-Moreau. Affect as a psychological primitive. *Advances in Experimental Social Psychology*, 41:167–218, 2009.

[9] M. S. Bartlett, G. Donato, J. R. Movellan, J. C. Hager, P. Ekman, and T. J. Sejnowski. Face image analysis for expression measurement and detection of deceit. In *Proceedings of the 6th Annual Joint Symposium on Neural Computation*, 1999.

[10] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 592–597. IEEE, 2004.

[11] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[12] J. N. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37(11):2049–2058, 1979.

[13] J. N. Bassili. Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2058, 1979.

[14] A. Beck, L. Canamero, and K. Bard. Towards an affect space for robots to display emotional body language. In *IEEE RO-MAN*, pages 464–469, 2010.

[15] A. Beck, L. Canamero, and K. Bard. Towards mapping emotive gait patterns from human to robot. In *IEEE RO-MAN*, pages 258–263, 2010.

[16] H. Beigi. *Fundamentals of Speaker Recognition*. Springer, 2011.

[17] P. Belin, S. Fillion-Bilodeau, and F. Gosselin. The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, 40(2):531–539, 2008.

[18] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14:585–591, 2001.

[19] P. Berens. CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software*, 31(10):1–21, 2009.

[20] E. Bevacqua, M. Mancini, R. Niewiadomski, and C. Pelachaud. An expressive ECA showing complex emotions. In *Proceedings of the AISB Annual Convention, Newcastle, UK*, pages 208–216, 2007.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Machine Learning Research*, 3:993–1022, 2003.

[22] J. Block. Studies in the phenomenology of emotions. *The Journal of Abnormal and Social Psychology*, 54(3):358–363, 1957.

[23] S. H. Blumberg and C. E. Izard. Patterns of emotion experiences as predictors of facial expressions of emotion. *Merrill-Palmer Quarterly*, pages 183–197, 1991.

[24] B. Braathen, M. S. Bartlett, G. Littlewort, E. Smith, and J. R. Movellan. An approach to automatic recognition of spontaneous facial actions. In *Proceedings*

Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pages 360–365. IEEE, 2002.

[25] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 4(42):335–359, 2008.

[26] G. Caridakis, K. Karpouzis, and S. Kollias. User and context adaptive neural networks for emotion recognition. *Neurocomputing*, 71(13-15):2553–2562, 2008.

[27] S.-C. Chang and J.-M. Jin. *Computation of Special Functions*. Wiley, 1996.

[28] Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006.

[29] L. S.-H. Chen. *Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction*. PhD thesis, Citeseer, 2000.

[30] S. Cherry. Anger management [emotion recognition]. *Spectrum, IEEE*, 42(4):16, 2005.

[31] C. Clavel, I. Vasilescu, L. Devillers, and T. Ehrette. Fiction database for emotion detection in abnormal situations. In *Proceedings of the International Conference on Spoken Language Processing*, 2004.

[32] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, and C. Sedogbo. The SAFE corpus: Illustrating extreme emotions in dynamic situations. In *The Workshop Programme Corpora for Research on Emotion and Affect*, page 76, 2006.

[33] J. Cohn and K. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(2):121–132, 2004.

[34] J. F. Cohn. Advances in behavioral science using automated facial image analysis and synthesis [social sciences]. *Signal Processing Magazine*, 27(6):128–133, 2010.

[35] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology*, 36(1):35–43, 1999.

[36] D. Collett and T. Lewis. Discriminating between the von mises and wrapped normal distributions. *Australian Journal of Statistics*, 23(1):73–79, 1981.

[37] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Computer Vision-ECCV*, pages 484–498. Springer, 1998.

[38] T. F. Cootes and C. J. Taylor. Combining point distribution models with shape models based on finite element analysis. *Image and Vision Computing*, 13(5):403–409, 1995.

[39] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Training models of shape from sets of examples. In *BMVC92*, pages 9–18. Springer, 1992.

[40] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[41] E. Costantini, F. Pianesi, and M. Prete. Recognising emotions in human and synthetic faces: The role of the upper and lower parts of the face. In *Proceedings*

of the 10th International Conference on Intelligent User Interfaces, pages 20–27. ACM, 2005.

[42] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz. What a neural net needs to know about emotion words. *Computational Intelligence and Applications*, pages 109–114, 1999.

[43] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. FEELTRACE: An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[44] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine*, 18(1):32–80, 2001.

[45] W. A. Cunningham and J. J. Van Bavel. Varieties of emotional experience: Differences in object or computation? *Emotion Review*, 1(1):56–57, 2009.

[46] W. A. Cunningham and P. D. Zelazo. The development of iterative reprocessing. *Developmental Social Cognitive Neuroscience*, page 81, 2010.

[47] W. A. Cunningham, P. D. Zelazo, et al. Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences*, 11(3):97–104, 2007.

[48] M. Dahmane and J. Meunier. Continuous emotion recognition using gabor energy filters. In *Affective Computing and Intelligent Interaction*, pages 351–358. Springer, 2011.

[49] A. R. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harvest Books, 1999.

[50] C. Darwin. *The Expression of the Emotions in Man and Animals*. Oxford University Press, USA, 1998.

[51] R. J. Davidson. Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & Emotion*, 12(3):307–330, 1998.

[52] R. J. Davidson, K. R. Scherer, and H. H. Goldsmith. *Handbook of Affective Sciences*. Oxford University Press New York, 2003.

[53] J. Davitz. *The Language of Emotion*. Personality and Psychopathology. Academic Press, 1969.

[54] E. Dellandrea, N. Liu, and L. Chen. Classification of affective semantics in images based on discrete and dimensional models of emotions. In *International Workshop on Content-Based Multimedia Indexing (CBMI, 2010)*, pages 1–6. IEEE, 2010.

[55] S. Demirbuga, E. Sahin, I. Ozver, S. Aliustaoglu, E. Kandemir, M. D. Varkal, M. Emul, and H. Ince. Facial emotion recognition in patients with violent schizophrenia. *Schizophrenia Research*, 144(1):142–145, 2013.

[56] R. Descartes. *The Passions of the Soul*. Hackett Publishing Company, 1989.

[57] L. Devillers, L. Vidrascu, and L. Lamel. 2005 special issue:Cchallenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

[58] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[59] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1):33–60, 2003.

[60] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen. The sensitive artificial listner: An induction technique for generating emotionally coloured conversation. In *LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko*, pages 1–4. ELRA, 2008.

[61] E. Douglas-Cowie, R. Cowie, and M. Schröder. A new emotion database: Considerations, sources and scope. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[62] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, et al. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Affective Computing and Intelligent Interaction*, pages 488–500. Springer, 2007.

[63] E. Douglas-Cowie and WP5-members. List of databases presented by the HUMAINE association. http://emotion-research.net/wiki/Databases, May 2004.

[64] T. D. Downs and A. L. Gould. Some relationships between the normal and von mises distributions. *Biometrika*, 54(3-4):684–687, 1967.

[65] P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[66] P. Ekman. *Emotion in the Human Face*. Cambridge University Press, 2 edition, 1982.

[67] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[68] P. Ekman. Basic emotions. *Handbook of cognition and emotion*, 4:5–60, 1999.

[69] P. Ekman. Sixteen enjoyable emotions. *Emotion Researcher*, 18(2):6–7, 2003.

[70] P. Ekman. *Emotions Revealed: Recognising Faces and Feelings to Improve Communication and Emotional Life*. Holt Paperbacks, 2007.

[71] P. Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. WW Norton & Company, 2009.

[72] P. Ekman, R. J. Davidson, W. V. Friesen, et al. The duchenne smile: Emotional expression and brain physiology. *Journal of Personality and Social Psychology*, 58(2):342–353, 1990.

[73] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[74] P. Ekman and W. V. Friesen. *Unmasking the Face*. Englewood Cliffs, N. J. : Prentice-Hall, 1975.

[75] P. Ekman and W. V. Friesen. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1(1):56–75, 1976.

[76] P. Ekman and W. V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Actions*. Palo Alto: CA: Consulting Psychologists Press, 1978.

[77] P. Ekman, W. V. Friesen, P. Ellsworth, et al. *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. Pergamon Press New York, 1972.

[78] P. Ekman and E. L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[79] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien. Emotion on the road: Necessity, acceptance, and feasibility of affective computing in the car. *Advances in Human-Computer Interaction*, 2010.

[80] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien. Emotion on the road-Necessity, acceptance, and feasibility of affective computing in the car. *Advances in Human-Computer Interaction*, 2010, 2010.

[81] L. A. Feldman. Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, 69:153–153, 1995.

[82] N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1995.

[83] A. J. Fridlund. *Human Facial Expression: An Evolutionary View*. Academic Press, 1994.

[84] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407, 2000.

[85] N. H. Frijda. *The Emotions*. Cambridge University Press, 1986.

[86] S. Fukuda. *Emotional Engineering: Service Development*. Springer London, Limited, 2011.

[87] J. Gill and D. Hangartner. Circular data in political science and how to handle it. *Political Analysis*, 18(3):316–336, 2010.

[88] A. Graesser, B. McDaniel, P. Chipman, A. Witherspoon, S. D'Mello, and B. Gholson. Detection of emotions during learning with autotutor. In *Proceedings of the 28th Annual Meetings of the Cognitive Science Society*, pages 285–290. Citeseer, 2006.

[89] J. A. Gray. *The Neuropsychology of Anxiety*. Oxford University Press, 1 edition, 1982.

[90] M. Grimm, K. Kroschel, and S. Narayanan. The vera am mittag german audio-visual emotional speech database. In *IEEE International Conference on Multimedia and Expo, 2008*, pages 865–868. IEEE, 2008.

[91] B. Grundlehner, L. Brown, J. Penders, and B. Gyselinckx. The design and analysis of a real-time, continuous arousal monitor. In *Sixth International Conference on Workshop on Wearable and Implantable Body Sensor Networks*, pages 156–161, 2009.

[92] H. Gunes, M. A. Nicolaou, and M. Pantic. Continuous analysis of affect from voice and face. In *Computer Analysis of Human Behaviour*, pages 255–291. Springer, 2011.

[93] H. Gunes and M. Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Intelligent Virtual Agents*, pages 371–377. Springer, 2010.

[94] H. Gunes and M. Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International Conference on Pattern Recognition*, volume 1, pages 1148–1153. IEEE, 2006.

[95] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(1):64–84, 2009.

[96] H. Gunes, M. Piccardi, and M. Pantic. *From the Lab to the Real World: Affect Recognition using Multiple Cues and Modalities*. InTech Education and Publishing, 2008.

[97] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 2012.

[98] G. Hajcak and D. M. Olvet. The persistence of attention to emotion: Brain potentials during and after picture presentation. *Emotion*, 8(2):250, 2008.

[99] A. Hakim, S. Marsland, and H. W. Guesgen. A reliable hybrid technique for human face detection. In *VISAPP (2)*, pages 241–244, 2010.

[100] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[101] M. Heller and V. Haynal. Depression and suicide faces. *What the Face Reveals*, pages 398–407, 1997.

[102] D. R. R. O. D. Heylen and M. Poel. On the re-implementation of FEELTRACE in the NXT framework for emotion annotation in AMI. Technical report, University of Twente, The Netherlands, 2005.

[103] H. Hong, H. Neven, and C. V. der Malsburg. Online facial expression recognition based on personalized galleries. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 354–359. IEEE, 1998.

[104] C.-L. Huang and Y.-M. Huang. Facial expression recognition using model-based feature extraction and action parameters classification. *Journal of Visual Communication and Image Representation*, 8(3):278–290, 1997.

[105] T. S. Huang, M. A. Hasegawa-johnson, S. M. Chu, Z. Zeng, and H. Tang. Sensitive talking heads, 2009.

[106] I. Hupont, E. Cerezo, and S. Baldassarri. Sensing facial emotions in a continuous 2D affective space. In *IEEE International Conference on Systems Man and Cybernetics (SMC)*, pages 2045–2051, 2010.

[107] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. C. Karpouzis, and S. D. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Journal of Neural Networks*, 18:423–435, 2005.

[108] C. E. Izard. *The Face of Emotion*, volume 23. Appleton-Century-Crofts New York, 1971.

[109] C. E. Izard. *The Psychology of Emotions*. Springer, 1991.

[110] D. C. Jackson, C. J. Mueller, I. Dolski, K. M. Dalton, J. B. Nitschke, H. L. Urry, M. A. Rosenkranz, C. D. Ryff, B. H. Singer, and R. J. Davidson. Now you feel it, now you don't: Frontal brain electrical asymmetry and individual differences in emotion regulation. *Psychological Science*, 14(6):612–617, 2003.

[111] W. James. What is an emotion? *Mind*, 9(34):188–205, 1884.

[112] S. R. Jammalamadaka and A. Sengupta. *Topics in Circular Statistics*, volume 5. World Scientific Publishing Company, 2001.

[113] Q. Ji, P. Lan, and C. Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on Systems, Man and Cybernetics*, 36(5):862–875, 2006.

[114] P. N. Johnson-Laird and K. Oatley. Basic emotions, rationality, and folk theory. *Cognition & Emotion*, 6(3-4):201–223, 1992.

[115] I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 2002.

[116] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53. IEEE, 2000.

[117] I. Kanluan, M. Grimm, and K. Kroschel. Audio-visual emotion recognition using an emotion space concept. In *16th European Signal Processing Conference, Lausanne, Switzerland*, 2008.

[118] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007.

[119] M. M. Karnaze. A constructivist approach to defining human emotion: From George Kelly to Rue Cromwell. *Journal of Constructivist Psychology*, 26(3):194–201, 2013.

[120] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In *Artifical Intelligence for Human Computing*, pages 91–112. Springer, 2007.

[121] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[122] M. Kipp. ANVIL - A generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, 2001.

[123] T. Kirkland and W. A. Cunningham. Neural basis of affect and emotion. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(6):656–665, 2011.

[124] T. Kirkland and W. A. Cunningham. Mapping emotions through time: How affective trajectories inform the language of emotion. *Emotion-APA*, 12(2):268, 2012.

[125] N. H. Kuiper. Tests concerning random points on a circle. *Koninklijke Nederlandse Akademie van Wetenschappen, The Netherlands*, 1960.

[126] D. Kulic and E. A. Croft. Affective state estimation for human-robot interaction. *IEEE Transactions on Robotics*, 23(5):991–1000, 2007.

[127] P. J. Lang. Behavioral treatment and bio-behavioral assessment: Computer applications. *Technology in Mental Health Care Delivery Systems*, pages 119–137, 1980.

[128] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (iaps): Technical manual and affective ratings, 1999.

[129] G. Laurans, P. Desmet, and P. Hekkert. The emotion slider: A self-report device for the continuous measurement of emotion. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*, pages 1–6. IEEE, 2009.

[130] R. S. Lazarus. *Emotion and Adaptation*. Oxford University Press New York, 1991.

[131] R. S. Lazarus. Psychological stress in the workplace. *Fifty Years of the Research and Theory by R. S. Lazarus: An Analysis of Historical and Perennial Issues*, page 312, 1998.

[132] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *10th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1983–1986, 2009.

[133] H.-J. Lin, S.-H. Yen, J.-P. Yeh, and M.-J. Lin. Face detection based on skin color segmentation and svm classification. In *Second International Conference on Secure System Integration and Reliability Improvement*, pages 230–231. IEEE, 2008.

[134] G. C. Littlewort, M. S. Bartlett, and K. Lee. Faces of pain: Automated measurement of spontaneousa facial expressions of genuine and posed pain. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 15–21. ACM, 2007.

[135] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Journal of Image and Vision Computing*, 27(12):1741–1844, 2009.

[136] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 94–101. IEEE, 2010.

[137] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, April 1936.

[138] M. H. Mahoor, S. Cadavid, D. S.Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 1 & 2, pages 833–839, 2009.

[139] C. Z. Malatesta and J. M. Haviland. Learning display rules: The socialization of emotion expression in infancy. *Child Development*, pages 991–1003, 1982.

[140] K. V. Mardia. *Statistics of Directional Data*, volume 5. Academic Press London, 1972.

[141] S. Marsland. *Machine Learning: An Algorithmic Perspective*. Chapman and Hall CRC, 2009.

[142] A. Martinez and S. Du. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *The Journal of Machine Learning Research*, 13:1589–1608, 2012.

[143] P. Martins and J. Batista. Identity and expression recognition on low dimensional manifolds. In *IEEE International Conference on Image Processing*, pages 3341–3344. IEEE, 2009.

[144] G. Matthews, M. Zeidner, and R. D. Roberts. *Emotional Intelligence: Science and Myth*. Cambridge, MA: MIT Press, 2002.

[145] W. McDougall. *An Introduction to Social Psychology*. J.W. Luce & Company, 1921.

[146] W. McDougall. *An Introduction to Social Psychology*. Courier Dover Publications, 2003.

[147] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *IEEE International Conference on Multimedia and Expo*, pages 1079–1084. IEEE, 2010.

[148] A. Mehrabian. Communication without words. *Psychology Today*, pages 51–52, 1968.

[149] A. Merla and G. L. Romani. Thermal signatures of emotional arousal: a functional infrared imaging study. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS, 2007)*, pages 247–249. IEEE, 2007.

[150] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198, 2012.

[151] P. Michel and R. Kaliouby. Real time facial expression recognition in video using support vector machines. In *IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 258–264, 2003.

[152] M. Mihelj, D. Novak, and M. Munih. Emotion-aware system for upper extremity rehabilitation. In *International Conference on Virtual Rehabilitation*, pages 160–165, 2009.

[153] A. A. Miranda, Y.-A. Le Borgne, and G. Bontempi. New routes from minimal approximation error to principal components. *Neural Processing Letters*, 27(3):197–207, 2008.

[154] J. Morriss, A. N. Taylor, E. B. Roesch, and C. M. van Reekum. Still feeling it: The time course of emotional recovery from an attentional perspective. *Frontiers in Human Neuroscience*, 7, 2013.

[155] J. R. Movellan. Tutorial on gabor filters. *Open Source Document*, 2002.

[156] O. H. Mowrer. *Learning Theory and Behaviour*. Wiley New York, 1960.

[157] B. R. Nhan and T. Chau. Classifying affective states using thermal infrared imaging of the human face. *IEEE Transactions on Biomedical Engineering*, 57(4):979–987, 2010.

[158] M. A. Nicolaou, H. Gunes, and M. Pantic. Output-associative RVM regression for dimensional and continuous emotion prediction. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 16–23. IEEE, 2011.

[159] A. O' Toole, J. Harms, S. Snow, D. Hurst, M. Pappas, J. Ayyad, and H. Abdi. A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):812–816, 2005.

[160] K. Oatley and P. N. Johnson-Laird. Towards a cognitive theory of emotions. *Cognition & Emotion*, 1(1):29–50, 1987.

[161] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[162] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1990.

[163] A. Ortony and T. J. Turner. What's basic about basic emotions. *Psychological Review*, 97(3):315–331, 1990.

[164] C. E. Osgood. Studies on the generality of affective meaning systems. *American Psychologist*, 17(1):10–28, 1962.

[165] I. Pandzic and R. Forchheimer. *MPEG-4 Facial Animation - The standard, Implementations and Applications*. John Wiley and Sons, 2002.

[166] J. Panksepp. Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences*, 5(03):407–422, 1982.

[167] J. Panksepp. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, USA, 1998.

[168] M. Pantic and L. J. M. Rothkrantz. An expert system for recognition of facial actions and their intensity. In *Seventeenth National Conference on Artificial Intelligence (AAAI-2001) / Twelfth Innovative Applications of Artificial Intelligence Conference*, pages 1026–1033, 2000.

[169] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 317–320, 2005.

[170] C. Peter, L. Axelrod, S. Afzal, H. Agius, E. Crane, and M. Balaam. Emotion in HCI-real world challenges. *Emotion in HCI Workshop*, 2009.

[171] R. W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

[172] J. C. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*. MIT press, 1999.

[173] R. Plutchik. Section of psychology: Outlines of a new theory of emotions. *Transactions of the New York Academy of Sciences*, 20(5 Series II):394–403, 1958.

[174] R. Plutchik. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, 1980.

[175] R. Plutchik. A general psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience. Theories of Emotion*, pages 3–33, 1980.

[176] R. Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553, 1982.

[177] R. Plutchik. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*. American Psychological Association (APA), 2002.

[178] R. E. Plutchik and H. R. Conte. *Circumplex Models of Personality and Emotions*. American Psychological Association, 1997.

[179] J. Prinz. Which emotions are basic? *Emotion, Evolution, and Rationality*, pages 69–87, 2004.

[180] B. Qian and K. Rasheed. Hurst exponent and financial market predictability. In *Proceedings of the 2nd IASTED International Conference on Financial Engineering and Applications*, pages 203–209, 2004.

[181] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis, and S. Kollias. Parameterized facial expression synthesis based on MPEG-4. *EURASIP Journal on Applied Signal Processing*, 2002(1):1021–1038, 2002.

[182] K. O. Regan. Emotion and e-learning. *Journal of Asynchronous Learning Networks*, 7(3):78–92, 2003.

[183] M. Rehm and M. Wissner. Gamble - A multiuser game with an embodied conversational agent. In *Entertainment Computing*, pages 180–191. Springer, 2005.

[184] R. Reisenzein. What is a definition of emotion? and are emotions mental-behavioral processes? *Social Science Information*, 46(3):424–428, 2007.

[185] T. Ritchie, J. J. Skowronski, J. Hartnett, B. Wells, and W. R. Walker. The fading affect bias in the context of emotion activation level, mood, and personal theories of emotion change. *Memory*, 17(4):428–444, 2009.

[186] G. I. Roisman, J. L. Tsai, and K.-H. S. Chiang. The emotional integration of childhood experience: physiological, facial expressive, and self-reported emotional response during the adult attachment interview. *Developmental Psychology*, 40(5):776–789, 2004.

[187] I. J. Roseman. Cognitive determinants of emotion: A structural theory. *Review of Personality & Social Psychology*, 1984.

[188] J. Rothwell, Z. Bandar, J. O'Shea, and D. McLean. Silent talker: A new computer-based system for the analysis of facial cues to deception. *Applied Cognitive Psychology*, 20(6):757–777, 2006.

[189] J. Russell. Studies on the generality of affective meaning systems. *Journal of Personality and Social Psychology*, 10(36):1152–1168, 1978.

[190] J. Russell. Measures of emotion. *Emotion: Theory, Research, and Experience*, 4:83–112, 1989.

[191] J. Russell. How shall an emotion be called? *Circumplex Models of Personality and Emotions*, pages 205–220, 1997.

[192] J. Russell and J. M. Fernández-Dols. What does a facial expression mean? *The Psychology of Facial Expression*, page 1, 1997.

[193] K. R. Scherer. Psychological models of emotion. *The Neuropsychology of Emotion*, 137:162, 2000.

[194] R. Schleicher, S. Sundaram, and J. Seebode. Assessing audio clips on affective and semantic level to improve general applicability. *Journal of the Acoustical Society of America*, 126(6):3156–67, 2009.

[195] H. Schlosberg. Three dimensions of emotion. *Psychological Review*, 61(2):81–88, 1954.

[196] K. L. Schmidt and J. F. Cohn. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology*, 116(S33):3–24, 2001.

[197] H. Scholsberg. A scale for the judgment of facial expressions. *Journal of Experimental Psychology*, 29(6):497–510, 1941.

[198] M. Schroder, E. Bevacqua, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, M. V. E. de Sevin, and M. Wöllmer. A demonstration of audiovisual sensitive artificial listeners. In *Affective Computing and Intelligent Interactions*, pages 263–264, 2009.

[199] N. Sebe, I. Cohen, T. Gevers, and T. Huang. Emotion recognition based on joint visual and audio cues. In *International Conference on Pattern Recognition*, pages 1136–1139, 2006.

[200] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.

[201] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[202] Y.-S. Shin. Facial expression recognition based on emotion dimensions on manifold learning. In *Computational Science-ICCS 2007*, pages 81–88. Springer, 2007.

[203] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41, 2012.

[204] D. Sontag and D. Roy. Complexity of inference in latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1008–1016, 2011.

[205] M. Stephens. Random walk on a circle. *Biometrika*, 50(3-4):385–390, 1963.

[206] K. Sun, J. Yu, Y. Huang, and X. Hu. An improved valence-arousal emotion space for video affective content representation and recognition. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 566–569. IEEE, 2009.

[207] D. Terzopoulos and K. Waters. Physically-based facial modelling, analysis, and animation. *The Journal of Visualization and Computer Animation*, 1(2):73–80, 1990.

[208] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[209] S. Tomkins and B. Karon. *Affect, Imagery, Consciousness: The Negative Affects*. Affect, Imagery, Consciousness. Springer Pub. Co., 1962.

[210] S. S. Tomkins. Affect theory. *Approaches to Emotion*, pages 163–195, 1984.

[211] T. C. Tsai, J. J. Chen, and W. C. Lo. Design and implementation of mobile personal emotion monitoring system. In *International Conference on Mobile Data Management*, pages 430–435, 2009.

[212] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591. IEEE, 1991.

[213] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 149–149. IEEE, 2006.

[214] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In *International Conference on Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70, 2010.

[215] M. F. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, pages 38–45. ACM, 2007.

[216] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 4, 2001.

[217] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 2010.

[218] G. S. Watson. Circular statistics in biology. *Technometrics*, 24(4):336–336, 1982.

[219] J. B. Watson. *Behaviorism*. Transaction Pub, 1924.

[220] B. Weiner and S. Graham. An attributional approach to emotional development. *Emotions, Cognition, and Behaviour*, pages 167–191, 1984.

[221] C. Whissell. *The Dictionary of Affect in Language*, volume 4. The Measurement of Emotions: Academic press, New York, 1989.

[222] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec. A dictionary of affect in language: Reliability, validity, and applications. *Perceptual and Motor Skills*, 62(3):875–888, 1986.

[223] C. M. Whissell. Pleasure and activation revisited: Dimensions underlying semantic responses to fiftyrandomly selected emotional words. *Perceptual and Motor Skills*, 53(3):871–874, 1981.

[224] J. Whitehill, Z. Serpell, A. Foster, Y.-C. Lin, B. Pearson, M. Bartlett, and J. Movellan. Towards an optimal affect-sensitive instructional system of cognitive skills. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 20–25. IEEE, 2011.

[225] P. Wik and A. Hjalmarsson. Embodied conversational agents in computer assisted language learning. *Speech Communication*, 51(10):1024–1037, 2009.

[226] L. Wiskott. *Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis*. Deutsch, 1995.

[227] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of the Annual Conference of the International Speech Communication Association (ISCA), Interspeech*, pages 597–600, 2008.

[228] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In *Proceedings of the Annual Conference*

of the International Speech Communication Association (ISCA), Interspeech, pages 2362–2365, 2010.

[229] M. Yeasin, B. Bullot, and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video. *IEEE Transactions on Multimedia*, 8(3):500–508, 2006.

[230] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3D facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.

[231] J. H. Zar. *Biostatistical Analysis*. Prentice Hall International, 3 edition, 1996.

[232] P. D. Zelazo and W. A. Cunningham. Executive function: Mechanisms underlying emotion regulation. *Handbook of Emotion Regulation*, pages 135–158, 2007.

[233] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[234] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. S. Huang. Audio-visual affect recognition in activation-evaluation space. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 4–7. IEEE, 2005.

[235] X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3D dynamic facial expression database. In *Proceedings of 10th IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

[236] G. Zhou, Y. Zhan, and J. Zhang. Facial expression recognition based on selective feature extraction. In *Sixth International Conference on Intelligent Systems Design and Applications*, volume 2, pages 412–417. IEEE, 2006.